THE ONR WORKSHOP ON DISCRETE STRUCTURES IN CLASSIFICATION

05 MAY – 06 MAY 1992

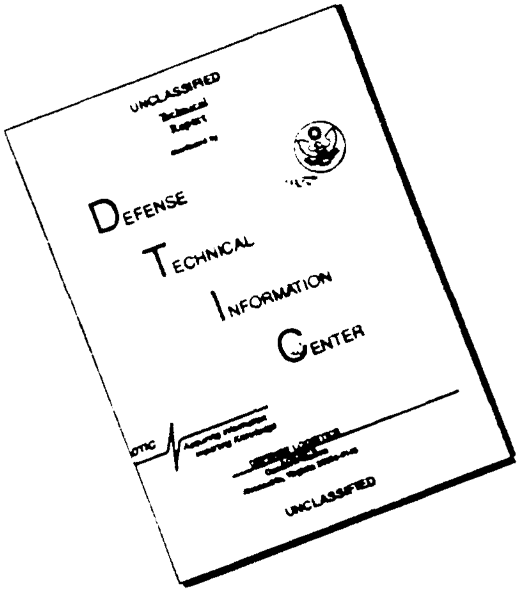
RAMADA RENAISSANCE HOTEL WASHINGTON-DULLES

PROCEEDINGS

Organized by: F.R. McMorris, University of Louisville
Marc J. Lipman, Office of Naval Research

GRANT: N00014-92-J-1574
R&T: 4111066---01

DISCLAIMER NOTICE



THIS DOCUMENT IS BEST QUALITY AVAILABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

PREAMBLE

The objective of this workshop was to determine the state-of-the-art in Discrete Structured Classification, and chart new directions of research relevant to Navy needs. Thus it had two goals. The first was for the Principal Investigators currently funded by ONR in this core area of Discrete Mathematics to inform each other and their Scientific Officer of their current research activities. The second was for the PIs and Navy scientists to meet each other and exchange problems and ideas with the hope of developing significant research partnerships.

The first goal was grandly achieved by direct PI presentations. Progress on the second was made through the problem session and roundtable discussions.

GROUP PICTURE
before the application of
CLASSIFICATION THEORY

ATTENDEES

Dr. W.O. Alltop
Code 3807
Applied Mathematics Group
Naval Air Weapons Center
China Lake, CA 93555-6001

Dr. Gary Chartrand
Department of Mathematics and Statistics
Western Michigan University
Kalamazoo, MI 49008

Dr. Donald W. Dearholt
Department of Computer Science
Mississippi State University
Mississippi State, MS 39762

Dr. Pierre Hansen
GERAD and École des Hautes études Commerciales
Département des Méthodes Quantitatives et Systèmes d'Information
5255 avenue Decelles
Montréal, Quebec H3T 1V6
Canada

Dr. M.F. Janowitz
Department of Mathematics and Statistics
University of Massachusetts
Amherst, MA 01003

Dr. Brigitte Jaumard
GERAD and École Polytechnique de Montréal
Département de Mathématiques Appliquées
Succursale A, Case Postale 6079
Montréal, Quebec H3C 3A7
Canada

Dr. Marc J. Lipman
Code 1111
Office of Naval Research
800 N. Quincy St.
Arlington, VA 22217

Dr. F.R. McMorris
Department of Mathematics
University of Louisville
Louisville, KY 40292

Dr. K.B. Reid
Department of Mathematics
California State University
San Marcos, CA 92069

Dr. Fred S. Roberts
Department of Mathematics
Rutgers University
New Brunswick, NJ 08903

Dr. Uriel G. Rothblum
RUTCOR
Rutgers University
New Brunswick, NJ 08903

Dr. Ralph Wachter
Code 1133
Office of Naval Research
800 N. Quincy St.
Arlington, VA 22217

ONR WORKSHOP ON
DISCRETE STRUCTURED CLASSIFICATION

Tuesday 5 May

0900	MARC LIPMAN
0930	BUCK McMORRIS
1030	PIERRE HANSEN BRIGITTE JAUMARD
1200	Lunch
1300	BROOKS REID
1400	DON DEARHOLT
1500	Break
1515	Problem session - MARC LIPMAN
1600	Roundtable
1800	Conference Dinner

Wednesday 6 May

0900	MARC LIPMAN
0930	GARY CHARTRAND
1030	break
1100	URIEL ROTHBLUM
1200	Lunch
1300	MEL JANOWITZ
1400	Roundtable
1500	Wrap

THE FOLLOWING PAGES CONSIST OF ABSTRACTS AND COPIES OF THE TRANSPARENCIES USED FOR THE PRESENTATIONS.

THE ONR PROGRAM IN DISCRETE STRUCTURED CLASSIFICATION

Dr. Marc J. Lipman
Scientific Officer for Discrete Mathematics
Office of Naval Research

Beginning in Fiscal Year 1992, The Office of Naval Research will have a long term program thrust in Discrete Structured Classification managed in the Discrete Mathematics Program. The goal of the thrust is to generate new methods for signal classification and robotic vision. This talk will present the basic structure and mathematical underpinnings of the new interest area, together with some specific problems of Navy interest and the mechanism for applying for support.

Dr. Marc J. Lipman, Code 1111
Office of Naval Research
800 N. Quincy St.
Arlington, VA 22217 - 5000
LIPMAN@OCNR-HQ.NAVY.MIL
703-696-4310

DISCRETE STRUCTURED CLASSIFICATION

22 FEB 1990
CODE 1111

DR. MARC J. LIPMAN, CODE 1111SP

A PROGRAM OF BASIC RESEARCH IN MODELING, CLUSTERING, AND CONSENSUS FORMATION IN SPACES OF DISCRETE STRUCTURES,

YIELDING NEW METHODS FOR SIGNAL CLASSIFICATION AND STEREO VISION SYNTHESIS.

22 FEB 1990
CODE 1111

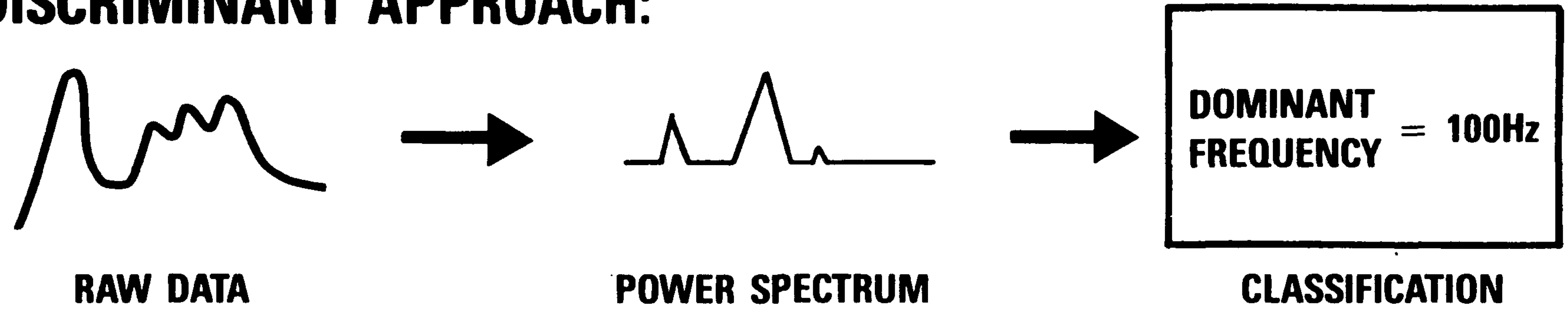
ISSUES

- CLUSTERING ON SETS OF DATA
  - GENERATION OF CANONICAL REPRESENTATIVES
- CLASSIFICATION IN NOISE, DISTORTION AND CHANGE OF ASPECT
  - DISTANCE TO CANONICAL REPRESENTATIVES
- RESOLUTION OF INCONSISTANT OBSERVATIONS
  - CONSENSUS FORMATION

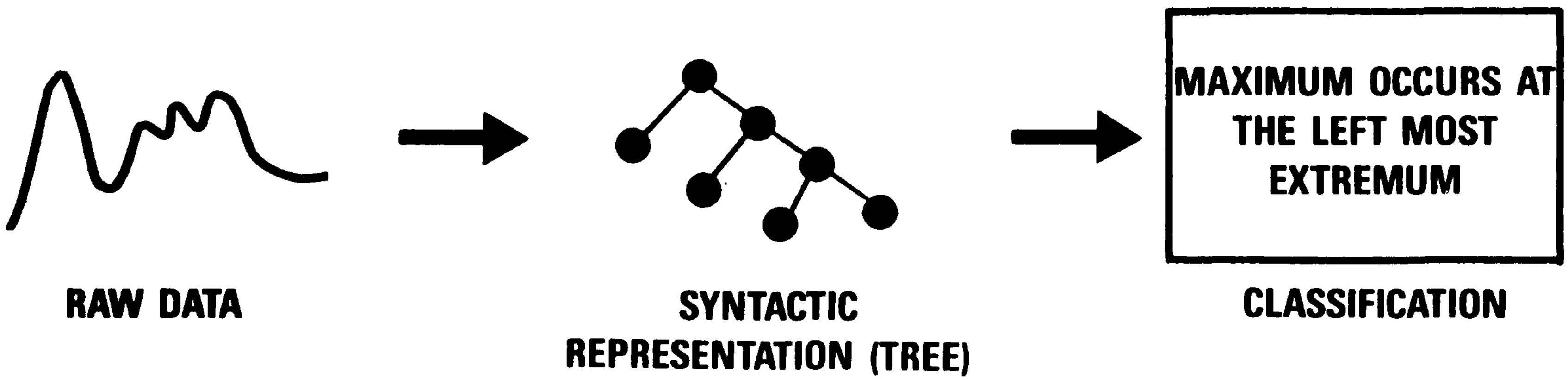
DISCRIMINANT vs SYNTACTIC CLASSIFICATION

22 FEB 1990
CODE 1111

DISCRIMINANT APPROACH:



SYNTACTIC APPROACH:



DISCRIMINANT vs. SYNTACTIC CLASSIFICATION

DISCRIMINANT APPROACH: (DECISION THEORETIC)

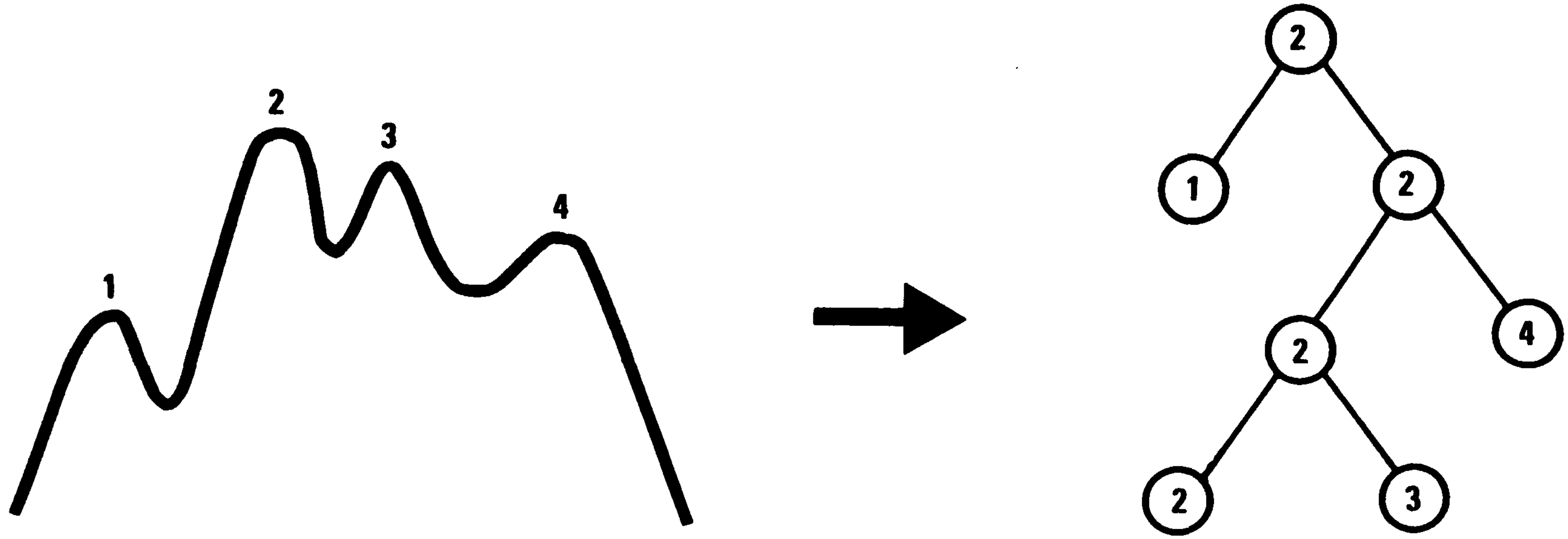
- EXTRACT FEATURES
- FORM VECTOR OF FEATURE VALUES
- CLUSTERS BASED ON FEATURE VECTOR DISCRIMINANTS

SYNTACTIC APPROACH: (STRUCTURAL)

- DESCRIBE DATA AS COMPOSITION OF COMPONENTS
- REPRESENT COMPONENTS WITH DISCRETE STRUCTURE
- CLUSTERS BASED ON DISCRETE METRICS

TREE STRUCTURED DATA – AN EXAMPLE

22 FEB 1990
CODE 1111



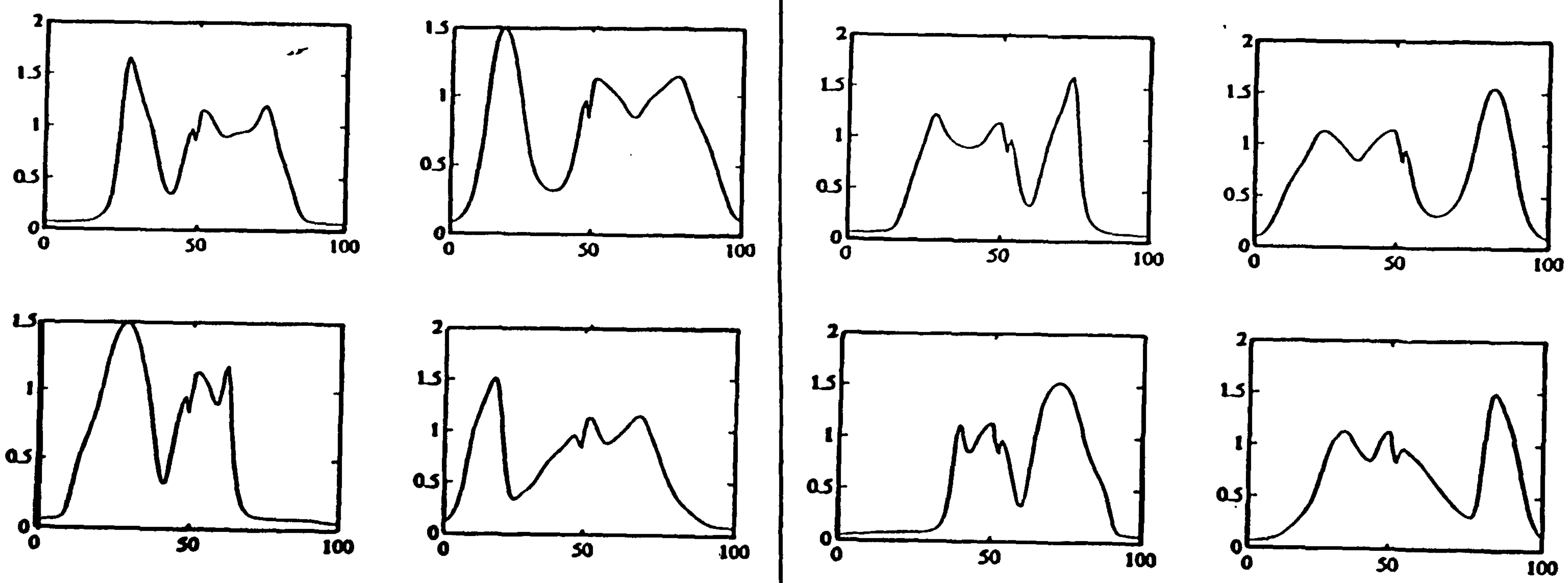
THE RELATIONAL TREE
CAPTURES STRUCTURAL FEATURES OF THE WAVEFORM

SYNTACTIC REPRESENTATION vs DISCRIMINANT METHODS

- ADVANTAGES OF SYNTACTIC REPRESENTATION
  - NO CENTERING PROBLEMS
  - NO DILATION/CONTRACTION PROBLEMS (ASPECT)
- STUMBLING BLOCK
  - NOISE
- SOLUTION
  - CONSENSUS FORMATION

CLUSTERING WAVEFORMS – AN EXAMPLE

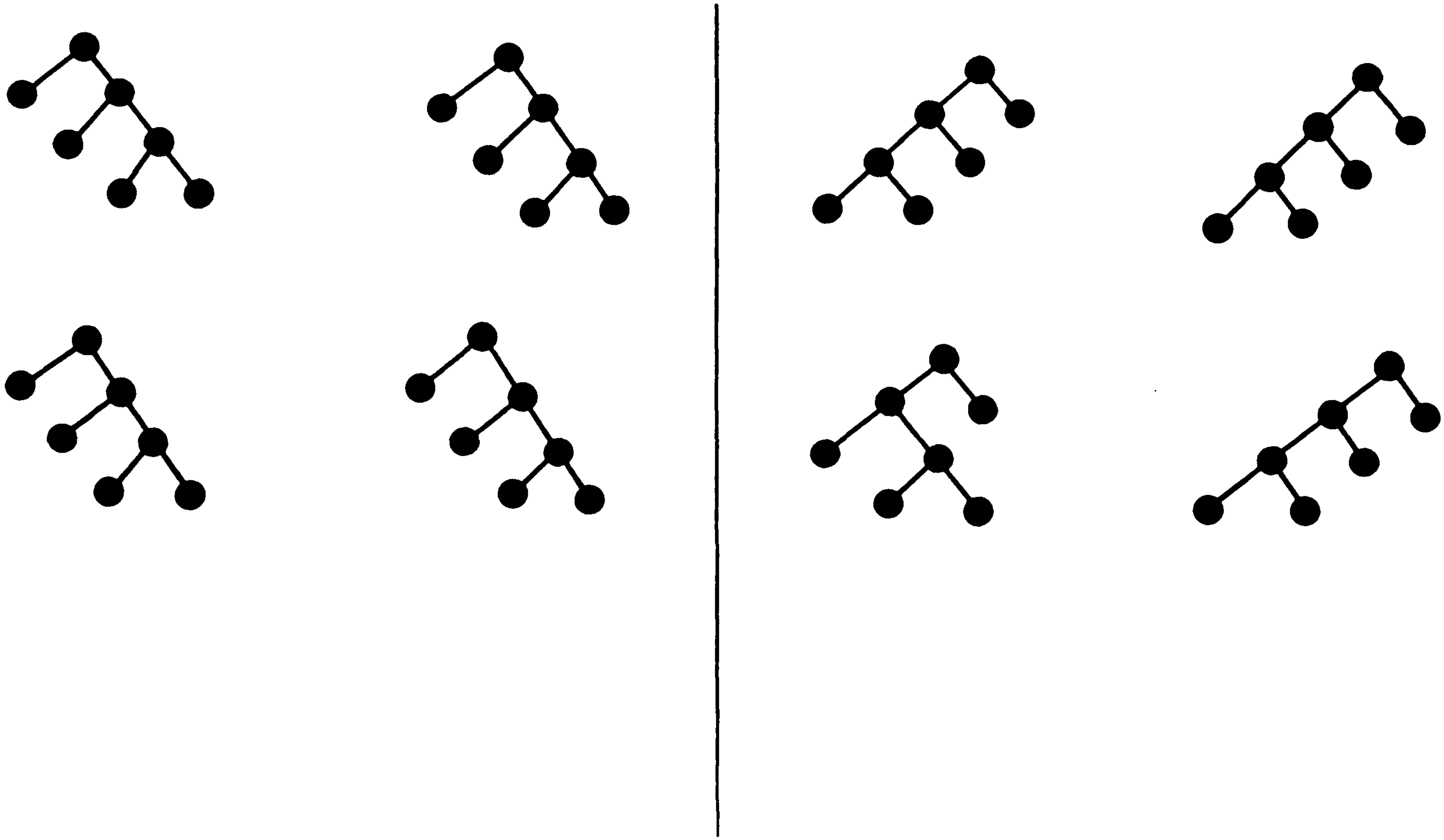
A SET OF 8 COMPUTER–GENERATED WAVEFORMS



CLUSTERING WAVEFORMS – AN EXAMPLE

22 FEB 1990
CODE 1111

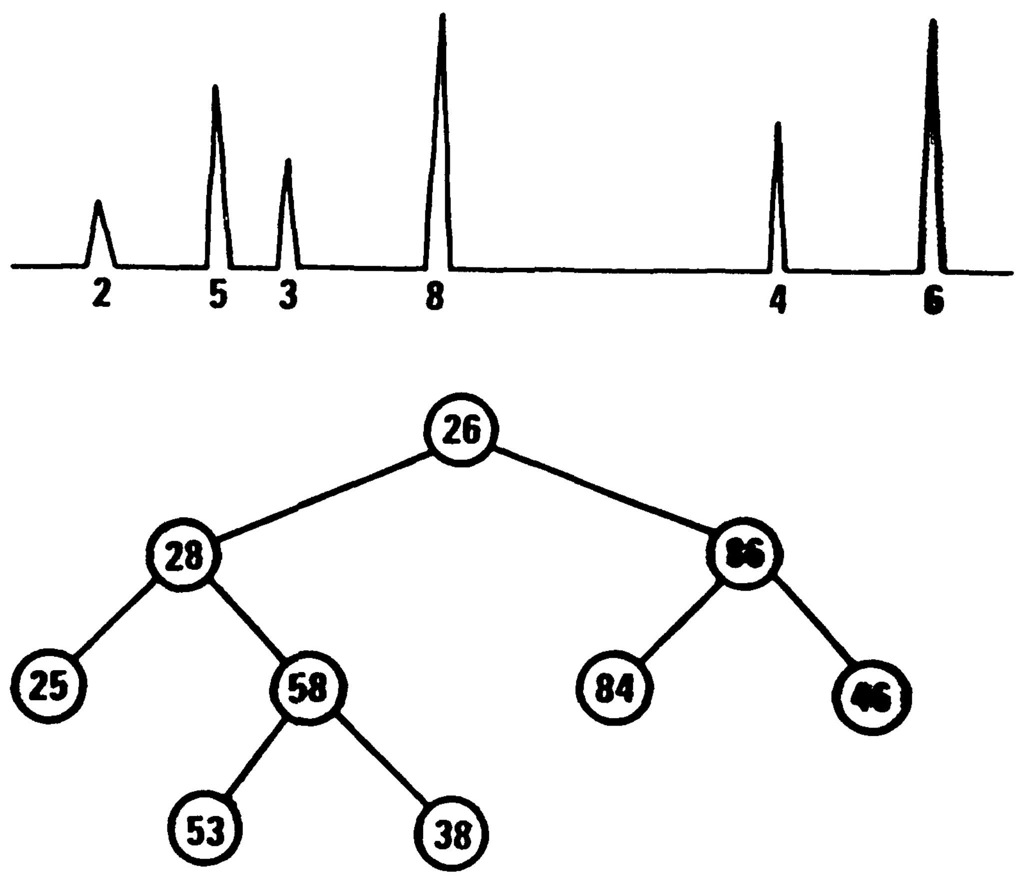
THE RELATIONAL TREES OF THESE WAVEFORMS:



TREE STRUCTURED DATA – B. T. LUCAS (NWC)

22 FEB 1990
CODE 1111

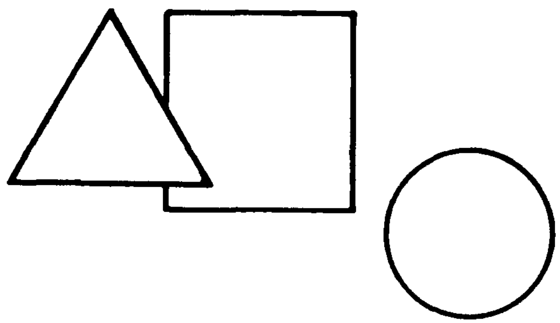
STRUCTURAL PROFILE DECOMPOSITION:



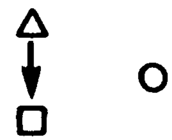
- NODES REPRESENT INTERVALS
- SPLITS DETERMINED BY THE MAXIMUM IN THE INTERVAL

PARTIALLY ORDERED SETS – AN EXAMPLE

22 FEB 1990
CODE 1111



OBSERVED SET

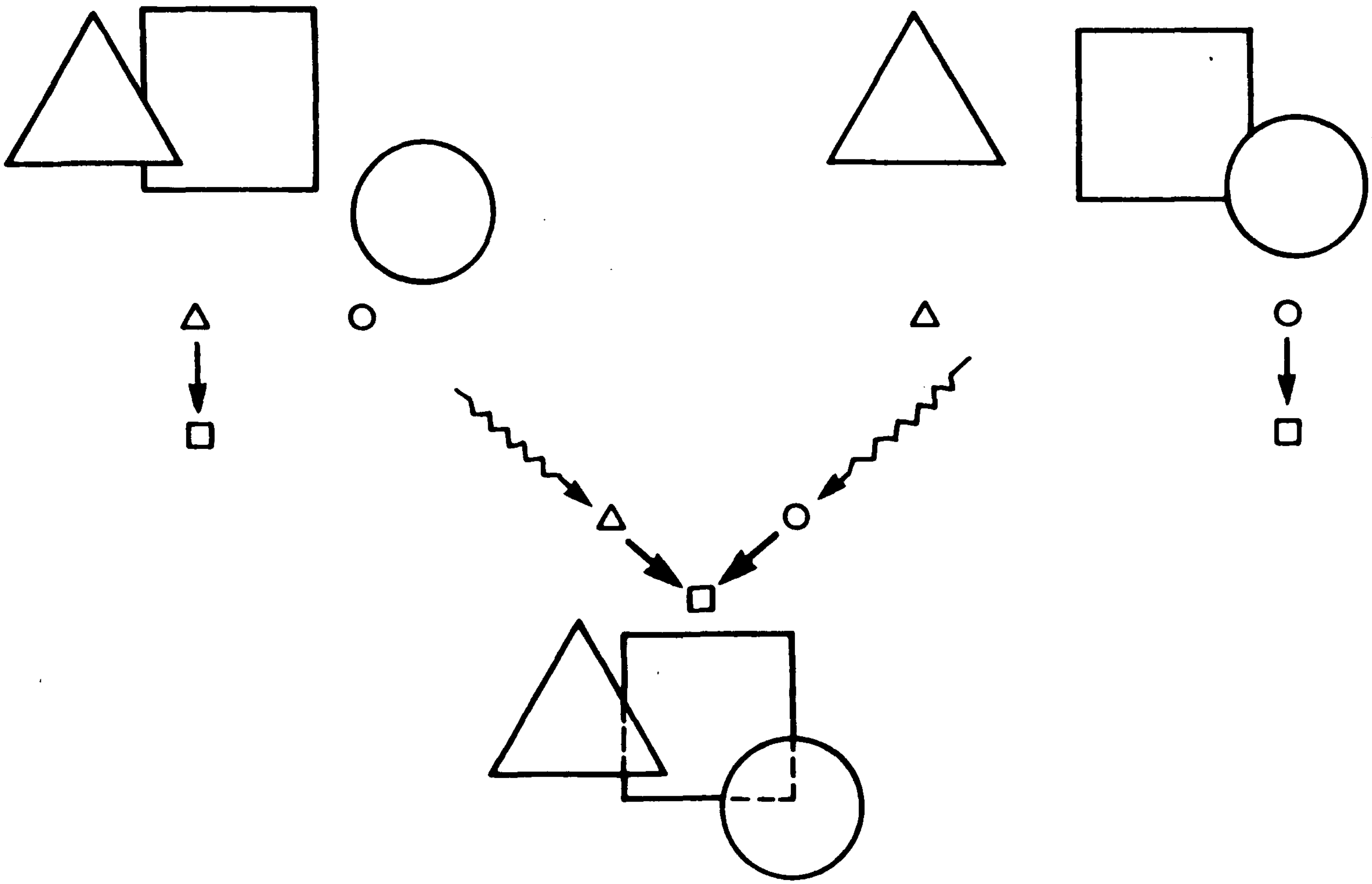


ASSOCIATED PARTIALLY ORDERED SET

CONSENSUS FORMATION — AN EXAMPLE
STEREOSCOPIC VISION

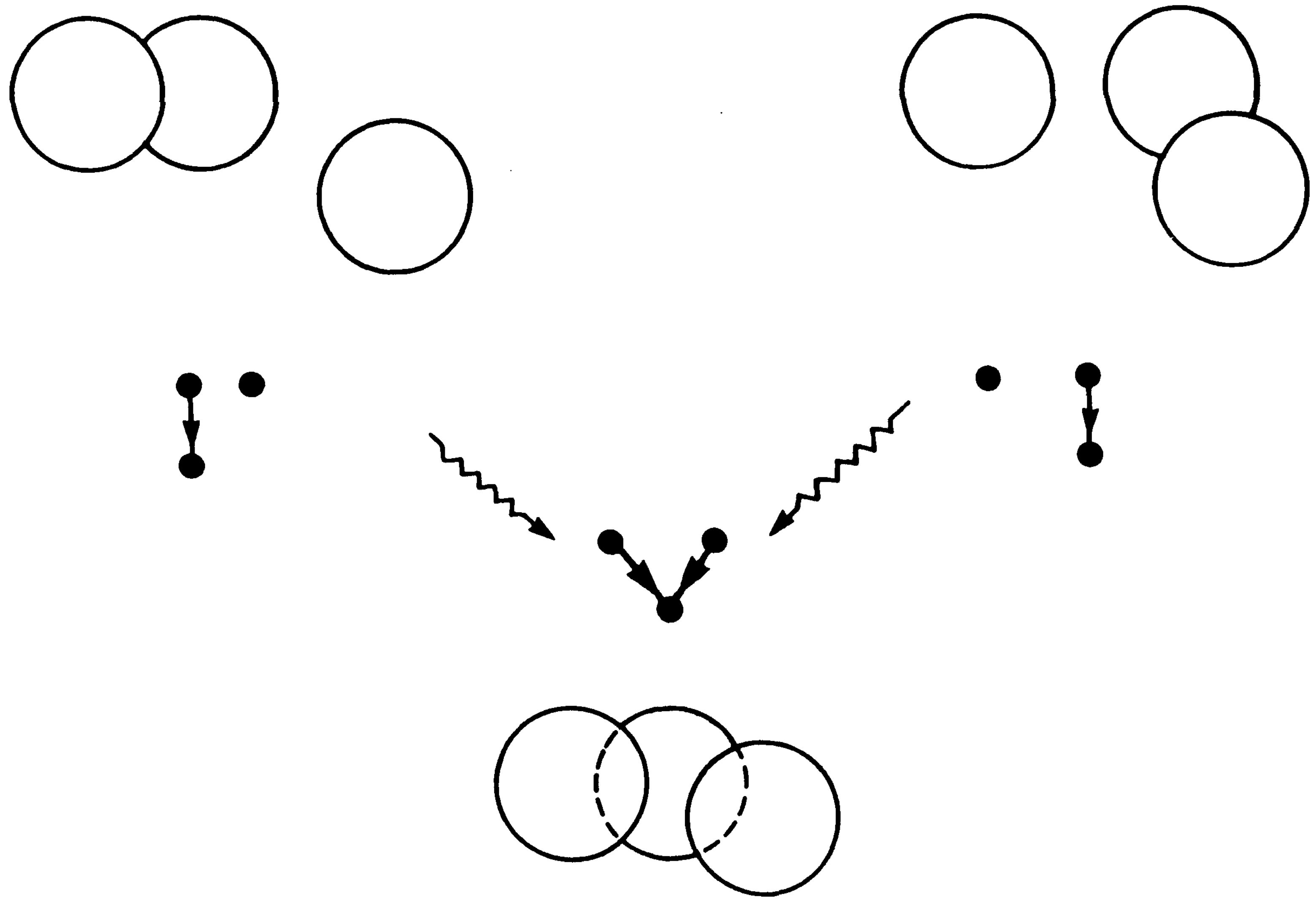
22 FEB 1990
CODE 1111

LEFT CAMERA: RIGHT CAMERA:



CONSENSUS FORMATION — AN EXAMPLE

22 FEB 1990
CODE 1111



MATHEMATICAL OBJECTIVES – NEAR TERM

22 FEB 1990
CODE 1111

- UNDERSTAND THE ROBUSTNESS (STABILITY) OF REPRESENTATIONS WITH RESPECT TO NOISE IN THE UNDERLYING DATA
- ALGORITHMIC ISSUES
  - COMPUTATIONAL TRACTABILITY
  - COMPUTATIONAL EFFICIENCY
- DEVELOP METRICS FOR CLUSTERING IN GENERAL SPACES OF DISCRETE STRUCTURES (BEYOND TREES)

MATHEMATICAL OBJECTIVES – LONG TERM

22 FEB 1990
CODE 1111

- DEVELOP A CONSENSUS ON CONSENSUS METHODS
  - MODE vs MEDIAN vs MEAN; UNION vs INTERSECTION
  - APPROPRIATENESS
  - CRITERIA BEYOND ROBUSTNESS
- DEVELOP A GENERAL "METRIC THEORY" FOR SPACES OF DISCRETE STRUCTURES
- PROGRESS FROM LABELED TO UNLABELED STRUCTURES

NAVY NEED

22 FEB 1990
CODE 1111

- AUTOMATED RECOGNITION OF ACOUSTIC AND RADAR SIGNATURES USING CLASSIFICATION IN DISCRETE METRIC SPACES
- MACHINE STEREO VISION FOR AUTONOMOUS VEHICLES

RESEARCH OBJECTIVES

* DEVELOPMENT OF METHODS FOR MODELING DIVERSE PHENOMENA WITH APPROPRIATE DISCRETE STRUCTURES

    * TREES
    * INTERVAL GRAPHS AND MULTIGRAPHS
    * INTERVAL ORDERS AND SEMIORDERS
    * DIGRAPHS
    * PARTIALLY ORDERED SETS

* UNDERSTANDING THE ROBUSTNESS OF DISCRETE MODELS WITH RESPECT TO NOISE IN THE UNDERLYING, UNCODED DATA

* DEVELOPMENT OF METRICS FOR SPACES OF DISCRETE STRUCTURES AND METHODS OF CONSENSUS FORMATION THAT EMPLOY THEM

RESEARCH ISSUES

* WHICH STRUCTURES BEST MODEL THE OBJECT OF INTEREST?

* DO THE DERIVED METRICS IN THE DISCRETE REPRESENTATIONS ACCURATELY REFLECT SIMILARITY/DISPARITY IN THE MODELED PHENOMENA?

* ARE THE TECHNIQUES OF MODELING AND SUBSEQUENT CLASSIFICATION AND INTERPRETATION COMPUTATIONALLY FEASIBLE?

An axiomatic approach to the aggregation of trees and other discrete structures

F.R. McMorris
Department of Mathematics
University of Louisville

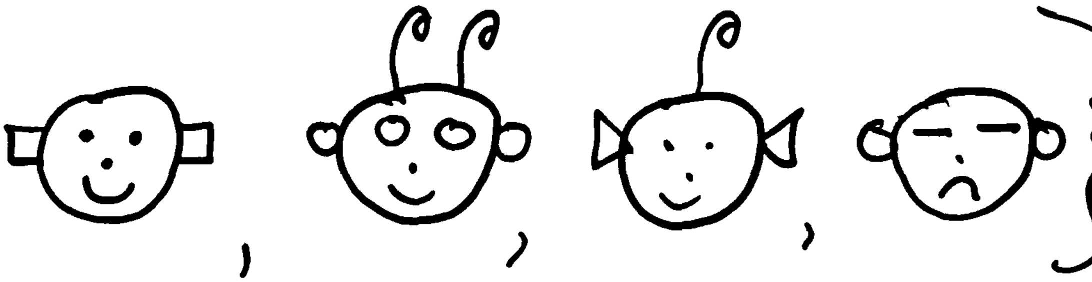
Suppose several classifications (usually structured as trees) have been constructed for the same set of objects. How can we (why should we) form a "consensus" classification that captures the common agreement of the original classifications? During the past twenty years many methods have been developed relevant to the comparison and consensus of classifications. One approach based on the central ideas of Social Choice Theory pioneered by K. Arrow, has been extended to produce a general axiomatic model to address the above question. This talk will present an overview of this approach and indicate some recent results.

An axiomatic approach to the aggregation of trees and other discrete structures

F.R. McMorris
UNIVERSITY OF LOUISVILLE

Example of a "typical" classification situation:

1. Given a set of objects S that we want to classify. (arrange into groups based on similarity)

$S = \{$  $\}$

2. Determine appropriate characters (features)

character	states
ear	□, ○, ◁
eye	•, ○, –
hair	None,

So, e.g.,  = (□, •, none)

3. Based on the character states, determine a basic data matrix D, or a dissimilarity coefficient

$$d : S \times S \to [0,1]$$

$$d(x,y) = d(y,x)$$

$$d(x,x) = 0.$$

4. Use a clustering algorithm on D to produce a classification for S.

Typical outputs are partitions (non-hierarchical classifications) and trees of various types.

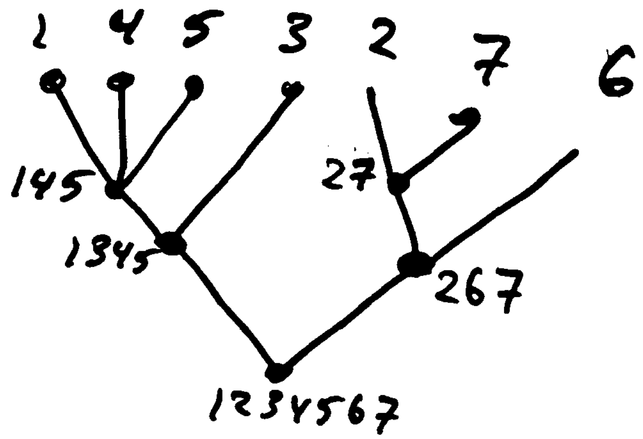
I will restrict attention to trees of the following type:

Let $S = \{1, \ldots, n\}$. An <u>n-tree</u> T is a subset of 2^S satisfying

1. $S \in T$, $\emptyset \notin T$
2. $\{i\} \in T \quad \forall i = 1, \ldots, n$
3. $A \cap B \in \{\emptyset, A, B\} \;\; \forall A, B \in T$

If $A \in T$, A is called a <u>cluster</u> of T.

Example:



Why consensus??

(i) Suppose we have k algorithms available to operate on D, the basic data matrix, and we use them all to produce trees $T_1, \ldots, T_k$.

What is the common-agreement tree $\mathbf{T}$?

$$(T_1, \ldots, T_k) \longmapsto T$$

(ii)

$$
\begin{tikzcd}
\text{Morphological characters used} \arrow[d] & & \text{Protein sequence characters used} \arrow[d] \\
T_1 \arrow[dr] & & T_2 \arrow[dl] \\
& T &
\end{tikzcd}
$$

– consensus tree used to indicated where the two approaches agree

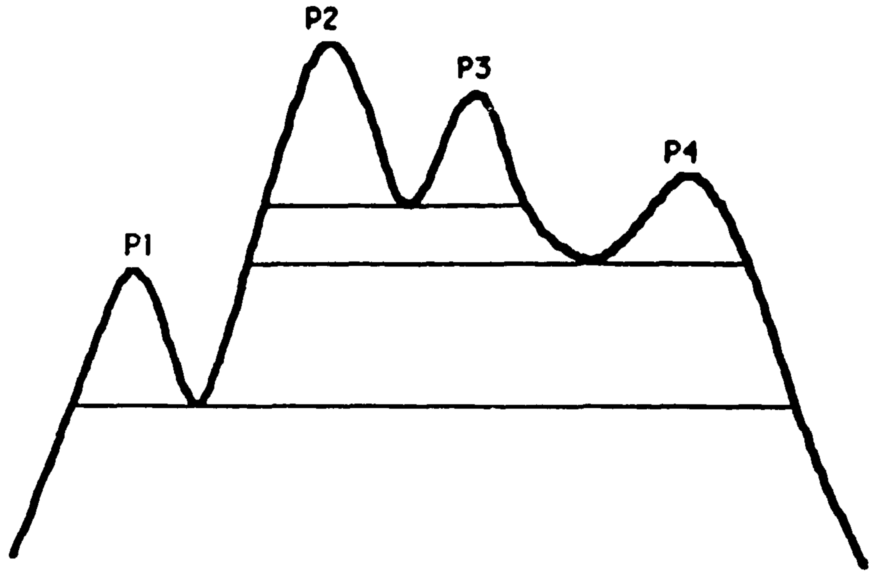


Figure 2.1 a
Waveform Segments

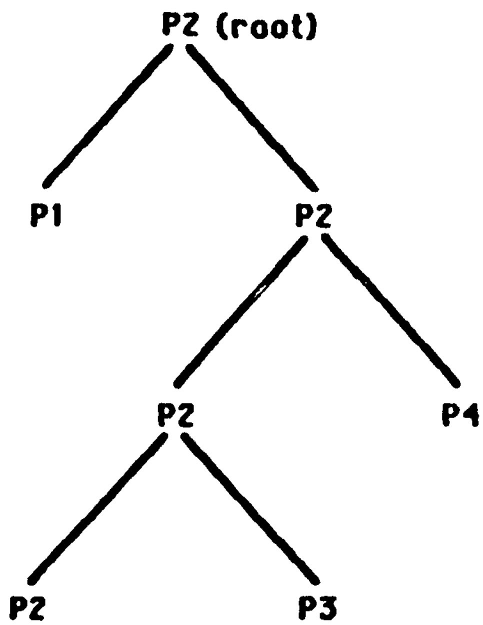


Figure 2.1 b
The Relational Tree

Erich and Foith enumerate several properties of these trees [1]. For our work, the most interesting is that RT's partition the set of one-dimensional functions into equivalence classes. The partitions may be viewed as clusters in a pattern feature

<u>The</u> <u>consensus</u> <u>problem</u> <u>for</u> <u>discrete</u> <u>structures</u>

Let $\mathcal{D}$ be the set of all discrete structures of a particular type. (graphs, digraphs, tree, partitions, ... satisfying certain conditions)

<u>The</u> <u>challenge</u>: Given k elements of $\mathcal{D}$, give a method for capturing the "common agreement" of these elements and representing this as another element of $\mathcal{D}$.

A <u>consensus</u> <u>function</u> is a map

$$C : \mathcal{D}^k \longrightarrow \mathcal{D}$$

$$c : \mathcal{D}^k \rightarrow 2^{\mathcal{D}} - \{\emptyset\} \qquad c : \cup \mathcal{D}^k \rightarrow 2^{\mathcal{D}} - \{\emptyset\}$$

$$C : \bigcup_{k \geq 1} \mathcal{D}^k \rightarrow \mathcal{D}$$

Two approaches to the consensus problem:

① The "armchair" approach.

② The "other" approach.

<u>Examples</u> <u>of</u> <u>the</u> <u>armchair</u> <u>method</u>:

What are "good" properties for $C: \mathcal{D}^k \to \mathcal{D}$?

Let $P = (D_1, \ldots, D_k) \in \mathcal{D}^k$ be a <u>profile</u> and suppose "' $x \in D_i$ "' $\forall i$ (x is a basic unit of info in $\mathcal{D}$).

Then require "' $x \in C(P)$ "'.

This is the <u>Pareto</u> <u>Condition</u>
(unanimity preserving)

Two approaches to the consensus problem:

① The "armchair" approach.

② The "other" approach.

<u>Examples of the armchair method</u>:

What are "good" properties for $C: \mathcal{D}^k \to \mathcal{D}$?

Let $P = (D_1, \ldots, D_k) \in \mathcal{D}^k$ be a <u>profile</u> and suppose "$x \in D_i$" $\forall i$ (x is a basic unit of info in $\mathcal{D}$).

Then require "$x \in C(P)$".

This is the <u>Pareto Condition</u> (unanimity preserving)

Research program based on the "Social choice" paradigm.

Generic independence conditions

Let $P = (D_1, \ldots, D_k)$ and $P' = (D_1', \ldots, D_k')$ be profiles and $X \subseteq S$. $D_1|_X = D_1'|_X \ldots$

Require: If $P|_X = P'|_X$, then

$$C(P)|_X = C(P')|_X .$$

For linear and weak orders on S, this is the classical "independence of irrelevant alternatives." (IIA)

Stability condition. $C: \bigcup_{k \geq 1} \mathbb{D}^k \to \mathbb{D}$

Require: For profiles P_1, P_2,

if $C(P_1) = C(P_2)$, then

$$C(P_1 P_1) = C(P_1) \quad (= C(P_2))$$

famous consequence of the armchair approach.

<u>Theorem</u> (K. Arrow, ~1957, ..., Nobel Prize Economics 1972)

Let $|S| > 2$ and $C : \mathcal{W}^k \to \mathcal{W}$ where $\mathcal{W}$ is the set of all weak orders on S.

If C satisfies the Pareto and IIA conditions, then C is dictatorial.

(i.e. $\exists\, i \ni \forall x, y \in S$; x strictly preferred to y by voter i implies x is strictly preferred to y in the consensus order.) Usually phrased as an "impossibility" result.

We have established various "impossibility" results for several classes of trees.

The power of independence.

Theorem (?, folklore) (WILSON, etc.)

Let $|S| > 2$ and $C : \mathcal{L}^k \to \mathcal{L}$ where $\mathcal{L}$ is the set of all linear orders on S. If C satisfies IIA, then C is constant, dictatorial, or persecutive.

Corollary: Arrow's theorem for $\mathcal{L}$.

$\exists j$ s.t. $(x,y) \in L_j \Rightarrow (y,x) \in C(i)$

For n-trees $\mathcal{T}$, there are many possible versions of "independence":

In a forthcoming paper in *Descrete Appl. Math.*, R.C. Powers and I proved that for ~~graph~~ graph theoretic trees ~~with no vertices of degree two~~, and all vertices of degree 1 assigned a label from S; the only consensus functions satisfying ~~[illegible]~~ "independence" are

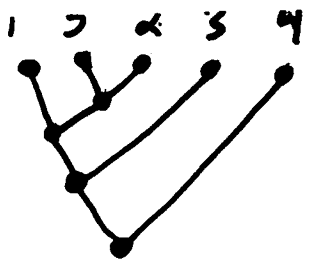
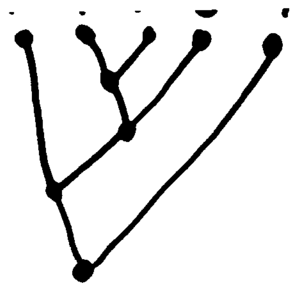
constant maps or *projections*!!

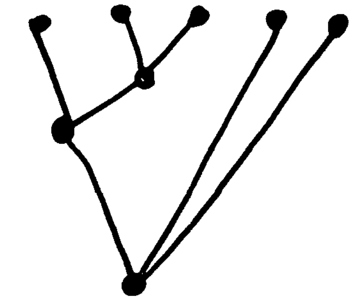
A reasonable new axiom and a strange consequence.

Notation:

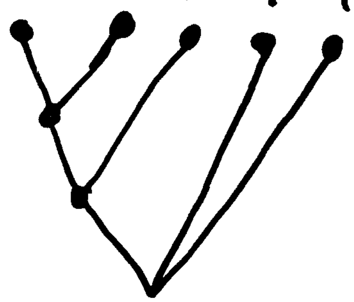
Let T be an n-tree with non-trivial clusters $A_1, \ldots, A_k$ and let $X \subseteq S$.

$T/_X$ will denote the n-tree whose non-trivial clusters are the non-trivial sets in $A_1 \cap X, \ldots, A_k \cap X$.

$T =$  $T' =$ 

$T\Big|_{\{1,2,5\}} =$ 

$= T'\Big|_{\{1,2,5\}}$

$T\Big|_{\{2,3,5\}} =$ 

$= T'\Big|_{\{2,3,5\}}$

Note that $\{1,2,5\} \in T$

and $\{2,3,5\} \in T'$

Consider this axiom

(I) For $C : \mathcal{T}^k \to \mathcal{T}$.

If $P, P' \in \mathcal{T}^k$ with

$P|_{\mathbb{X}} = P'|_{\mathbb{X}}$, then

$\mathbb{X} \in C(P)$ iff $\mathbb{X} \in C(P')$

<u>Proposition</u>: If C satisfies (I), then C does not satisfy the Pareto condition. (for $P \in \mathcal{T}^k$, $A \in T_i\ \forall i$ implies $A \in C(P)$.)

[i.e. (I) + Pareto is <u>really</u> impossible!!

and also (I) + "dictator" is impossible.

<u>Proof.</u> Assume C satisfies (I) and Pareto.

Let $P = (T, \ldots, T)$ and $P' = (T', \ldots, T')$ where T and T' are as before.

Since $$P\Big|_{\{1,2,5\}} = P'\Big|_{\{1,2,5\}}$$ and $\{1,2,5\} \in T$, Pareto condition gives $\{1,2,5\} \in C(P)$ and (I) gives $\{1,2,5\} \in C(P')$.

But $\{2,3,5\} \in T'$ so Pareto gives $\{2,3,5\} \in C(P')$ which contradicts treeness. //

Another generic independence condition, but based on the "units of information" x.

For $C: \mathcal{D}^k \to \mathcal{D}$,

$P = (D_1, \ldots, D_k)$, $P' = (D_1', \ldots, D_k')$

If $\{i : x \in D_i\} = \{i : x \in D_i'\}$,

then $x \in C(P)$ iff $x \in C(P')$

Recently Barthélemy, Janowitz, Leclerc and Monjardet have been developing an elegant general theory to handle "units of information" axioms in the context of semilattices.

Specific example follows:

Often $\mathcal{D}$ forms a nice type of semilattice like, for example, a median semilattice.

This is the case when $\mathcal{D} = \mathcal{T}_n$, the set of n-trees.

Barthélemy, Leclerc and Monjardet identified the join irreducibles in the semilattice $(\mathcal{D}, \leq)$ with "units of information in general.

In this case ~~the~~ $(\mathcal{D} = \mathcal{T}_n)$ the join irreducibles "are" the clusters.

Also note that a consensus function on the semilattice $(\mathcal{D}, \leq)$ is just a k-ary operation.

Example of the "other" method.

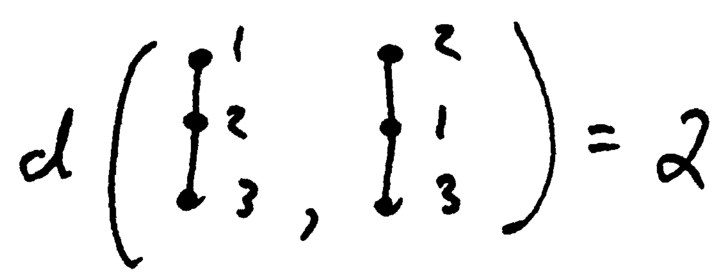
Let d be a metric on $\mathcal{D}$.

The median procedure is the consensus function $M: \bigcup_{k \geq 1} \mathcal{D}^k \to 2^{\mathcal{D}} \setminus \{\emptyset\}$ defined by

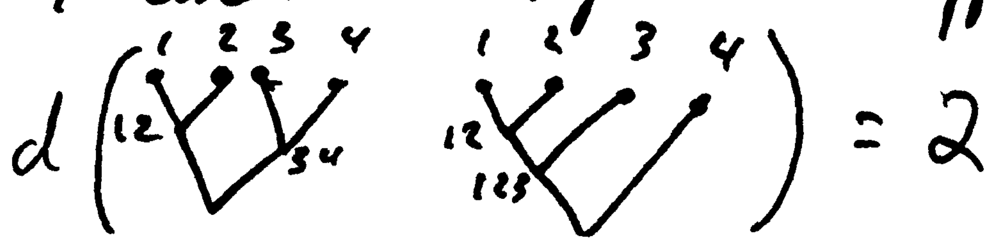
$$M(D_1, \ldots, D_k) = \{D \in \mathcal{D} : \sum_{i=1}^{k} d(D, D_i) \text{ is minimum}\}.$$

Two important cases:

(i) $\mathcal{D} = \mathcal{L}$ and d = symmetric difference metric



(ii) $\mathcal{D} = \mathcal{T}$ and d = symmetric difference metric



Computational complexity

(d) INSTANCE: $(\ell_1, \ldots, \ell_k) \in \mathcal{L}^k$ and a positive integer p.

QUESTION: Does there exist a linear order ℓ so that

$$\sum_{i=1}^{k} d(\ell, \ell_i) \le p \, ?$$

NP-complete

(Orlin ; Wakabayashi, 1986)

(b) INSTANCE: $(T_1, \ldots, T_k) \in \mathcal{T}^k$ and positive integer ρ.

QUESTION: Does there exist an n-tree T so that

$$\sum_{i=1}^{k} d(T, T_i) \leq \rho \text{ ?}$$

This is in P.

<u>Reason</u>: For $P = (T_1, \ldots, T_k)$ let Maj P $= \{A : A$ is a cluster in more than $1/2$ of the T_i's$\}$

Then Maj $P \in \mathcal{T}$ (!!) and is, in fact, an element of $M(P)$.

$\therefore$ only need check to see if

$$\sum d(\text{Maj } P, T_i) \leq \rho .$$

In fact we have

Theorem (Barthélemy, McMorris, 1986)

$M(P)$ consists of all the n-trees of the form

$Maj(P) \cup \{A_1, \ldots, A_m\}$ where, for $1 \le l \le m$, A_l is "compatible" with $Maj(P) \cup \{A_1, \ldots, A_{l-1}\}$ and A_i is a cluster in exactly $1/2$ of the n-trees in P.

Directions for future research:

(1) aggregation axioms for other discrete structures

(2) multiple DNA sequence alignment as a consensus problem W.H.E. DAY
(ARMCHAIR FIRST, PLEASE)

(3) greatest common subgraph derived axiomatically.

(4) discover new possibility results

O.N.R. WORKSHOP

MAY 5-6, 1992

K.B. Reid
California State Univ., San Marcos

- Agendas in majority voting
- Plurality preferences based on distances in graphs
- Centrality in graphs

SINCERE AND SOPHISTICATED DECISIONS VIA MAJORITY VOTING WITH AN AGENDA

- General problem: Where is the winning alternative under majority voting with an agenda?

- Here: – odd number of voters
  – each voter linearly orders the alternatives
  – majority rule used for each pairwise comparison of the alternatives

-

a_1 Admin. salaries

a_2 Faculty salaries

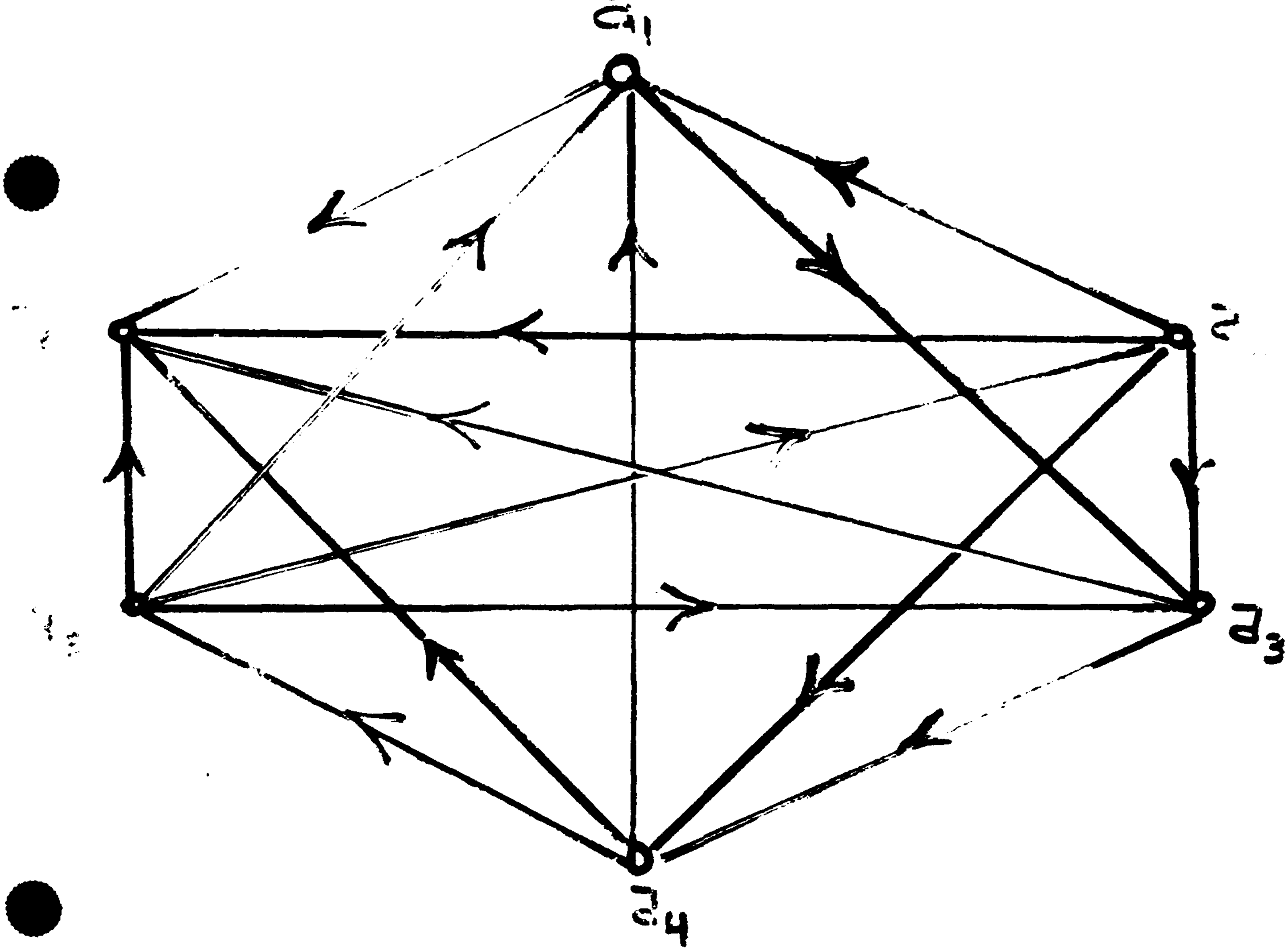
a_3 New gym

a_4 New music hall

a_5 New office bldg

a_6 New class bldg

V_1	V_2	V_3	V_4	V_5
[illegible]	a_1	a_3	a_6	[illegible]
a_5	a_4	a_2	a_5	a_4
a_1	a_5	a_4	a_2	a_1
a_3	a_2	a_5	a_4	a_6
a_4	a_6	a_1	a_1	[illegible]
[illegible]	a_3	[illegible]	[illegible]	[illegible]

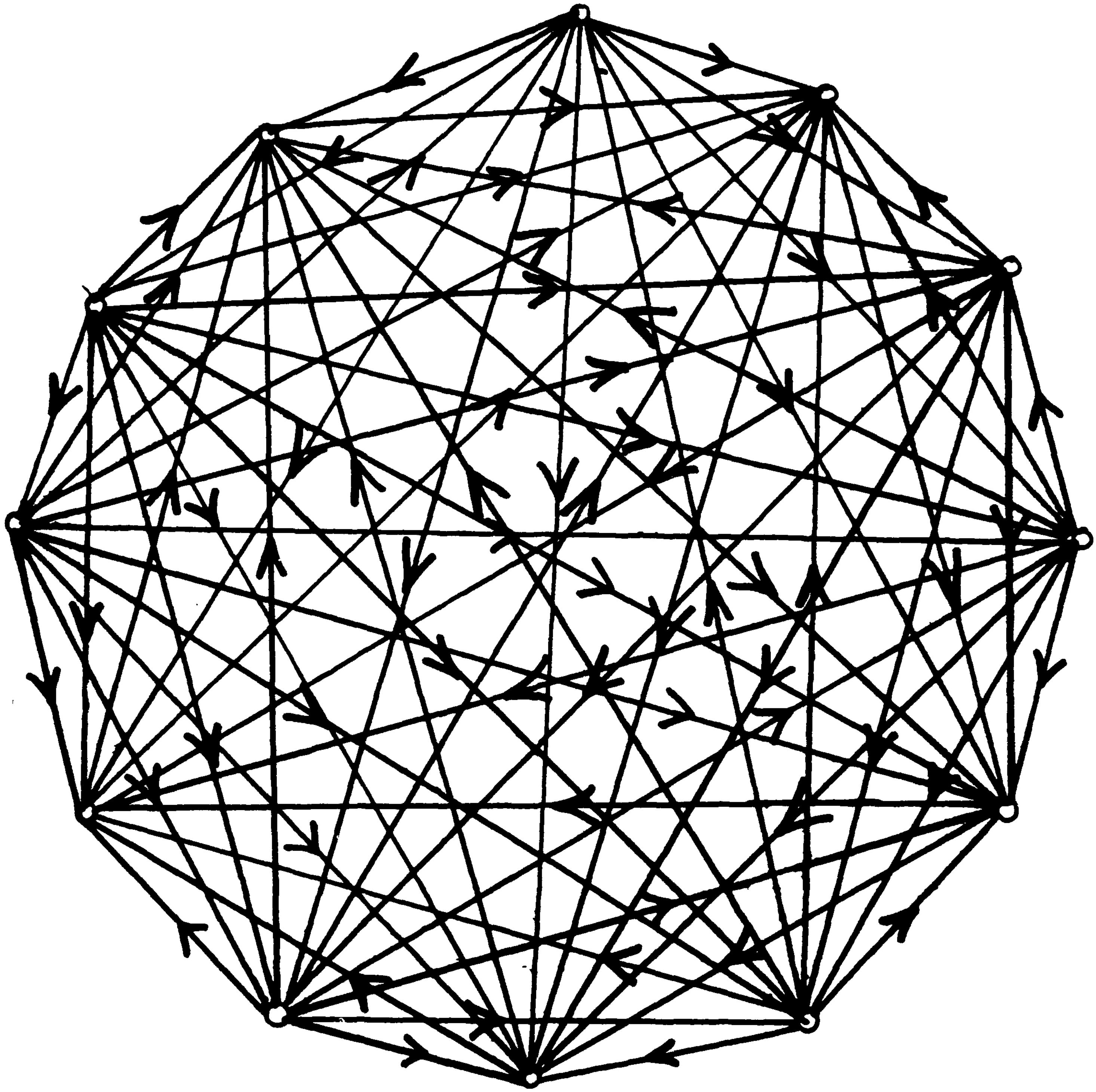


THEOREM

Every oriented graph arises as a majority preference digraph.

See J.W. Moon's 1968 monograph

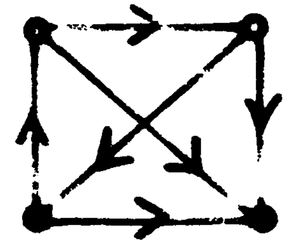
T... T...



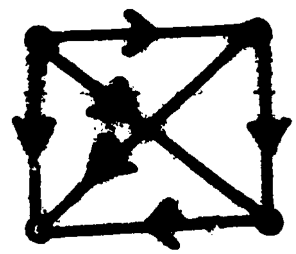
A tournament is an orientation of a complete graph.

THE FOUR 4-TOURNAMENTS

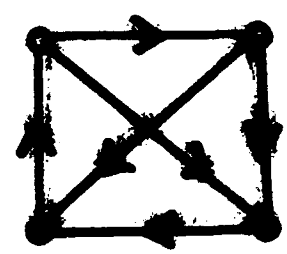
(0, 2, 2, 2)



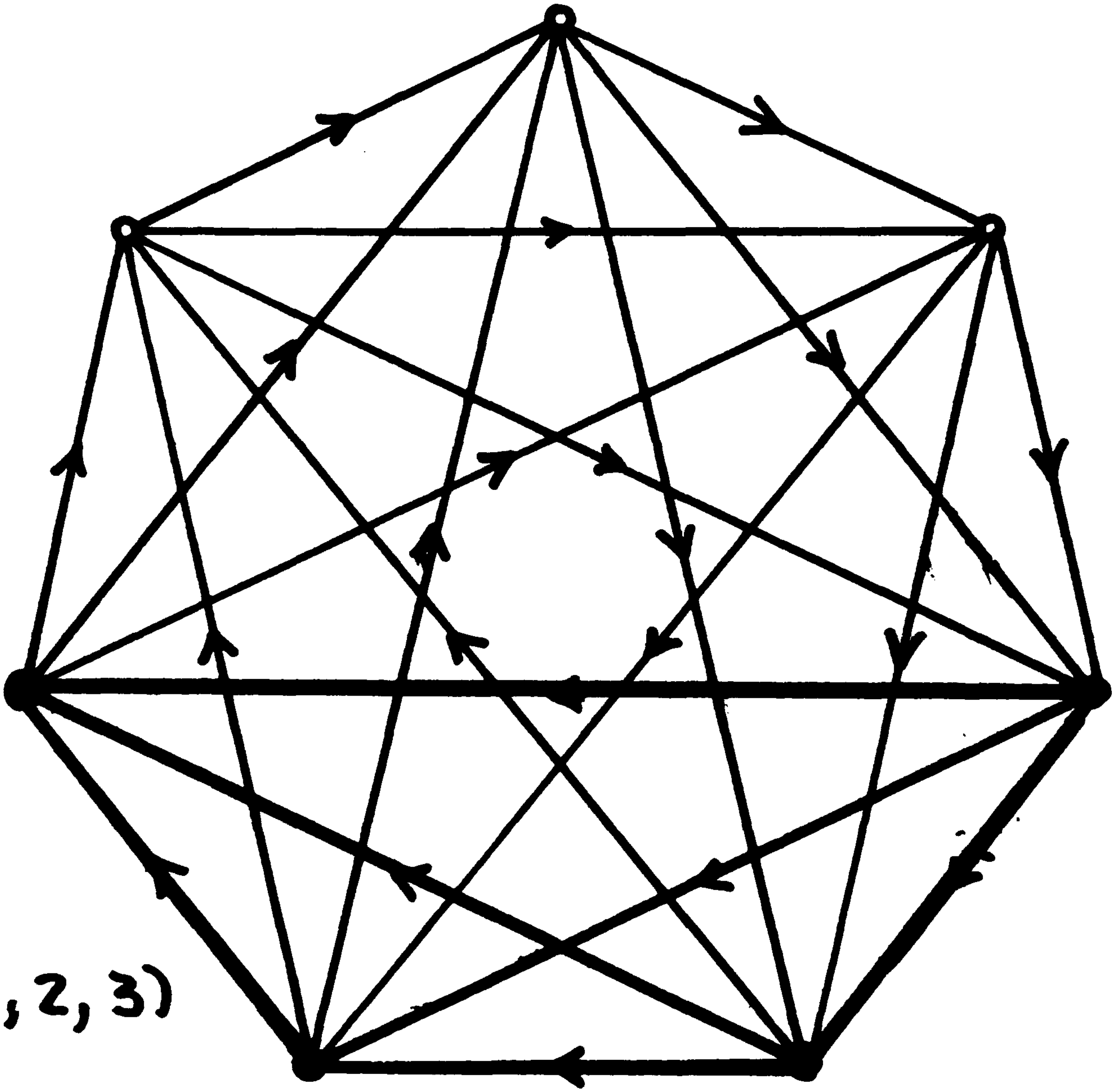
(1, 1, 1, 3)



(0, 1, 2, 3)



(1, 1, 2, 2)



(3, 3, 3, 3, 3, 3, 3)

TABLE A8

Tournaments

p	$T(p)$
1	1
2	1
3	2
4	4
5	12
6	56
7	456
8	6 880
9	191 536
10	9 733 056
11	903 753 248
12	154 108 311 168
13	48 542 114 686 912
14	28 401 423 719 122 304
15	31 021 002 160 355 166 848
16	63 530 415 842 308 265 100 288
17	244 912 778 438 520 759 443 245 824
18	1 783 398 846 284 777 975 419 600 287 232
19	24 605 641 171 260 376 770 598 003 978 281 472
20	645 022 068 557 873 570 931 850 526 424 042 500 096
21	32 207 364 031 661 175 384 456 332 280 036 660 040 346 624
22	3 070 169 883 150 468 336 193 188 889 176 239 554 269 865 953 280
23	559 879 382 429 394 075 397 997 876 821 117 309 031 348 506 639 435 776
24	1 956 920 276 575 218 760 843 168 426 608 334 827 851 734 377 775 365 039 898 624
25	131 326 696 677 895 002 131 450 257 709 457 767 557 170 027 052 967 027 982 788 816 896
26	169 484 335 125 246 268 100 514 597 385 576 342 667 201 246 238 506 672 327 765 919 863 947 264
27	421 255 599 848 131 447 082 003 884 098 323 929 861 369 544 621 589 389 269 735 693 986 231 100 612 608
28	2 019 284 625 667 208 265 086 928 694 043 799 677 058 780 746 074 756 618 649 807 453 554 008 410 636 526 845 952
29	18 691 296 182 213 712 407 784 892 577 100 643 237 772 159 079 535 345 610 331 272 616 359 410 643 727 554 822 061 146 112
30	334 493 774 260 141 796 028 606 267 674 709 437 232 608 940 215 918 926 763 659 414 050 175 507 824 571 200 950 884 097 540 096 000

TABLE A8

TOURNAMENTS

p	$T(p)$
1	1
2	1
3	2
4	4
5	12
6	56
7	456
8	6 880
9	191 536
10	9 733 056
11	903 753 248
12	154 108 311 168
13	48 542 114 686 912
14	28 401 423 719 122 304
15	31 021 002 160 355 166 848
16	63 530 415 842 308 265 100 288
17	244 912 778 438 520 759 443 245 824
18	1 783 398 846 284 777 975 419 600 287 232
19	24 605 641 171 260 376 770 598 003 978 281 472
20	645 022 068 557 873 570 931 850 526 424 042 500 096
21	32 207 364 031 661 175 384 456 332 260 036 660 040 346 624
22	3 070 169 883 150 468 336 193 188 889 176 239 554 269 865 953 280
23	559 879 382 429 394 075 397 997 876 821 117 309 031 348 506 639 435 776
24	1 956 920 276 575 218 760 843 168 426 608 334 827 851 734 377 775 365 039 898 624
25	131 326 696 677 895 002 131 450 257 709 457 767 557 170 027 052 967 027 982 788 816 896
26	169 484 335 125 246 268 100 514 597 385 576 342 667 201 246 238 506 672 327 765 919 863 947 264
27	421 255 599 848 131 447 082 003 884 098 323 929 861 369 544 621 589 389 269 735 693 986 231 100 612 608
28	2 019 284 625 667 208 265 086 928 694 043 799 677 058 780 746 074 756 618 649 807 453 554 008 410 636 526 845 952
29	18 691 296 182 213 712 407 784 892 577 100 643 237 772 159 079 535 345 610 331 272 616 359 410 643 727 554 822 061 146 112
30	334 493 774 260 141 796 028 606 267 674 709 437 232 608 940 215 918 926 763 659 414 050 175 507 824 571 200 950 884 097 540 096 000

$A = (a_1, a_2, \ldots, a_m)$ an agenda

SINCERE VOTING (UNDER AMENDMENT PROCEDURE)

Construct a "sincere sequence" $C = (c_1, c_2, \ldots, c_m)$

$$c_1 = a_1$$

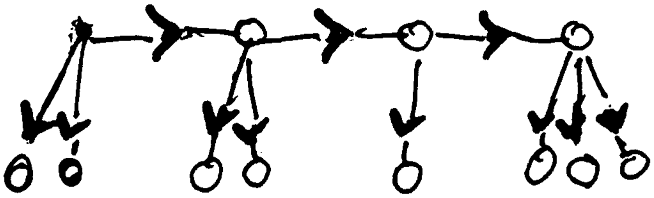
$$1 < i \leq m, \quad c_i = \begin{cases} a_i & \text{if } a_i \to c_{i-1} \\ c_{i-1}, & \text{if } c_{i-1} \to a_i \end{cases}$$

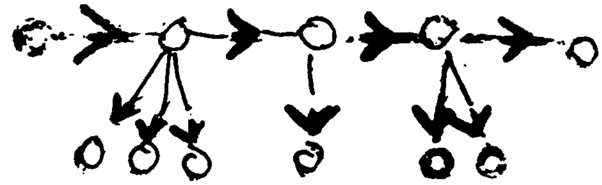
DEFINITION: c_m is the <u>sincere decision</u>

The $n-1$ votes in the sincere voting process correspond to $n-1$ arcs in T that form

- a rooted spanning tree, rooted at the sincere decision. This tree is the *decision tree* of T.

THEOREM. Let T be a tournament and let X denote a spanning subdigraph of T.
X is the decision tree of T relative to some agenda if and only if X is a "special orientati of a caterpillar.



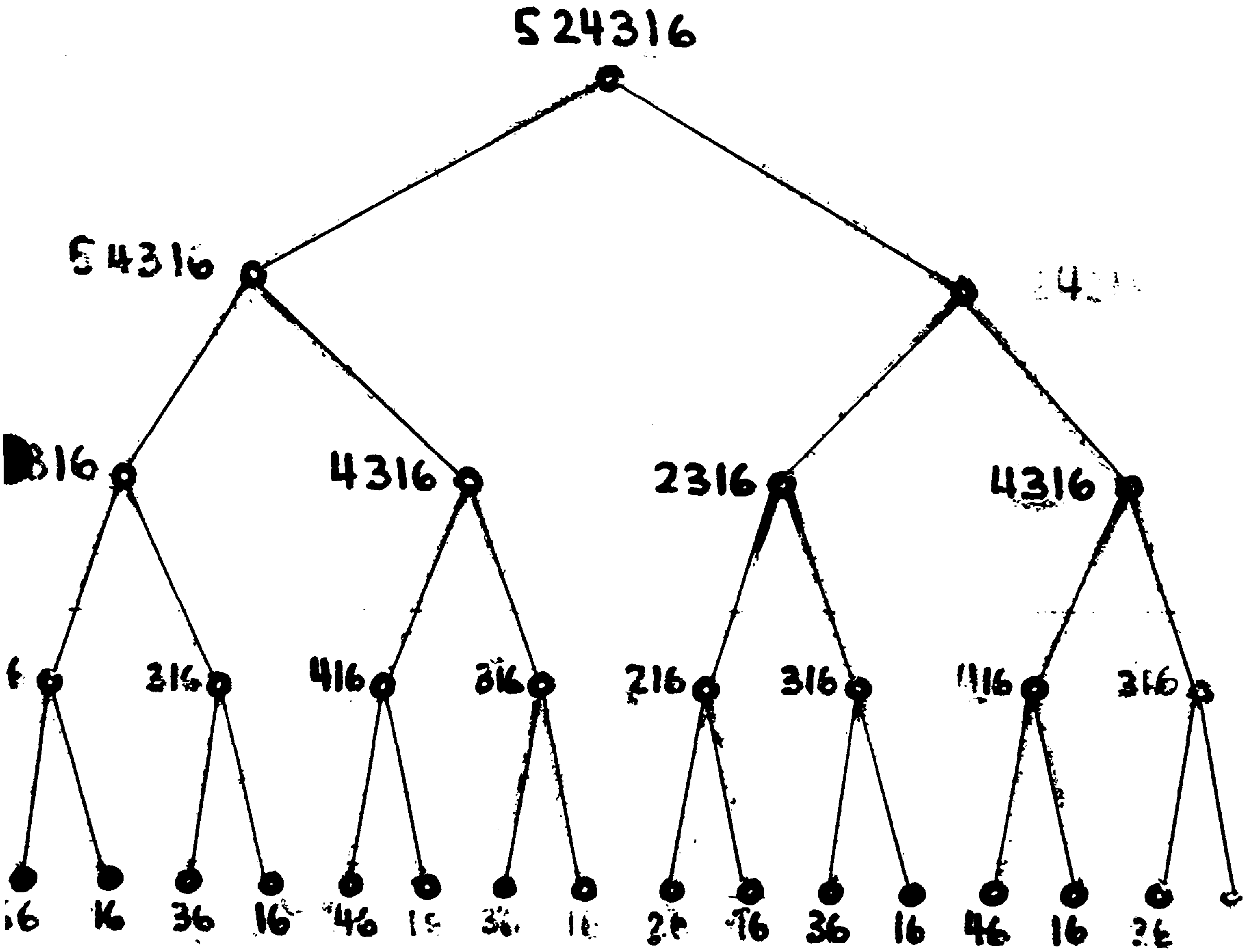


By the previous theorem the sincere decision under any agenda must lie in T^*.

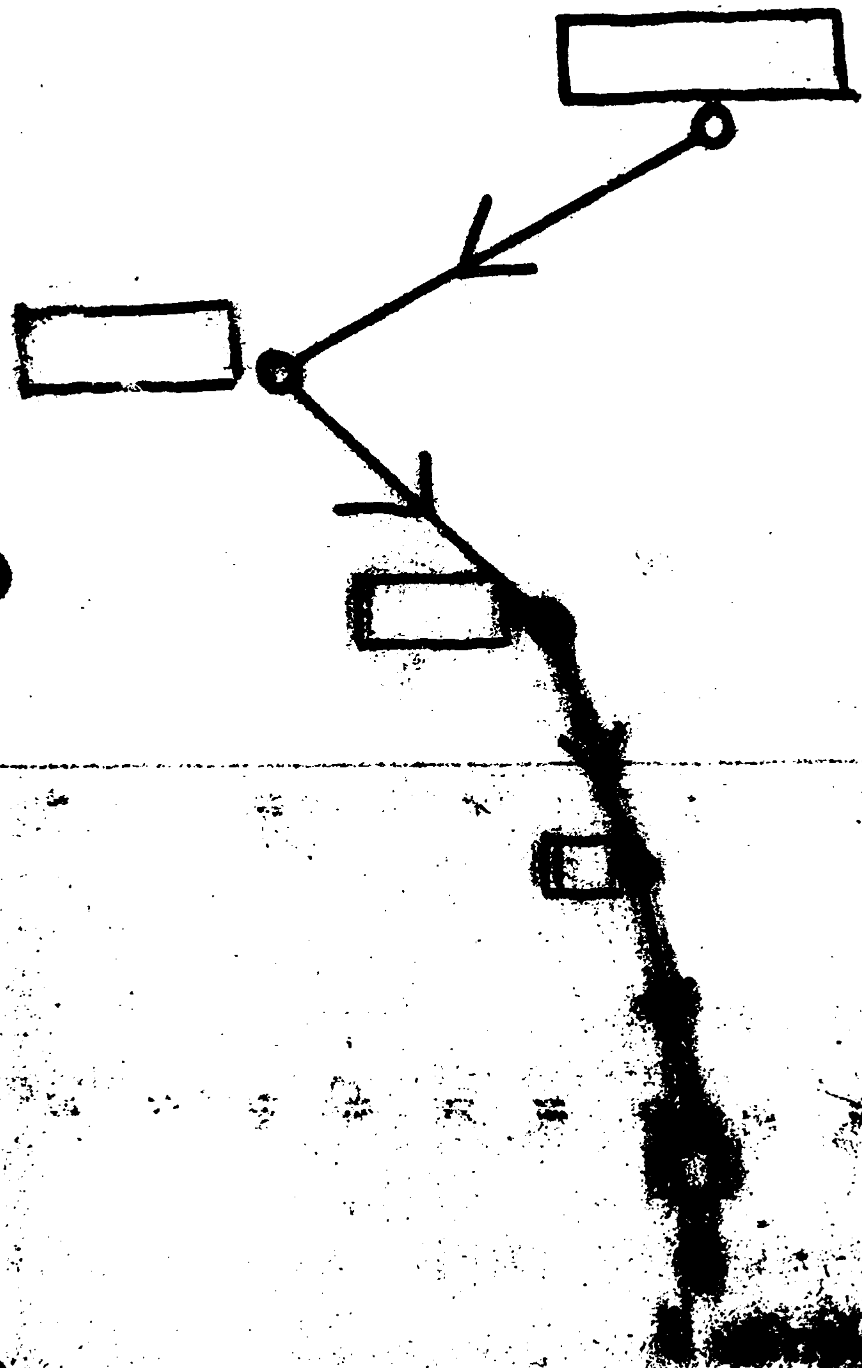
And, if $x \in T^*$ $\exists$ a spanning path in T starting at x, say $x \to y \to \cdots \to z$.

Then under agenda $A = (z, \ldots, y, x)$, x is the sincere decision.

524316 denotes $(a_5, a_2, a_4, a_3, a_1, a$



Sincere majority voting with agenda 524316



T a majority tournament on alternatives $\{a_1, a_2, \ldots, a_m\}$

$A = (a_1, a_2, \ldots, a_m)$ an agenda

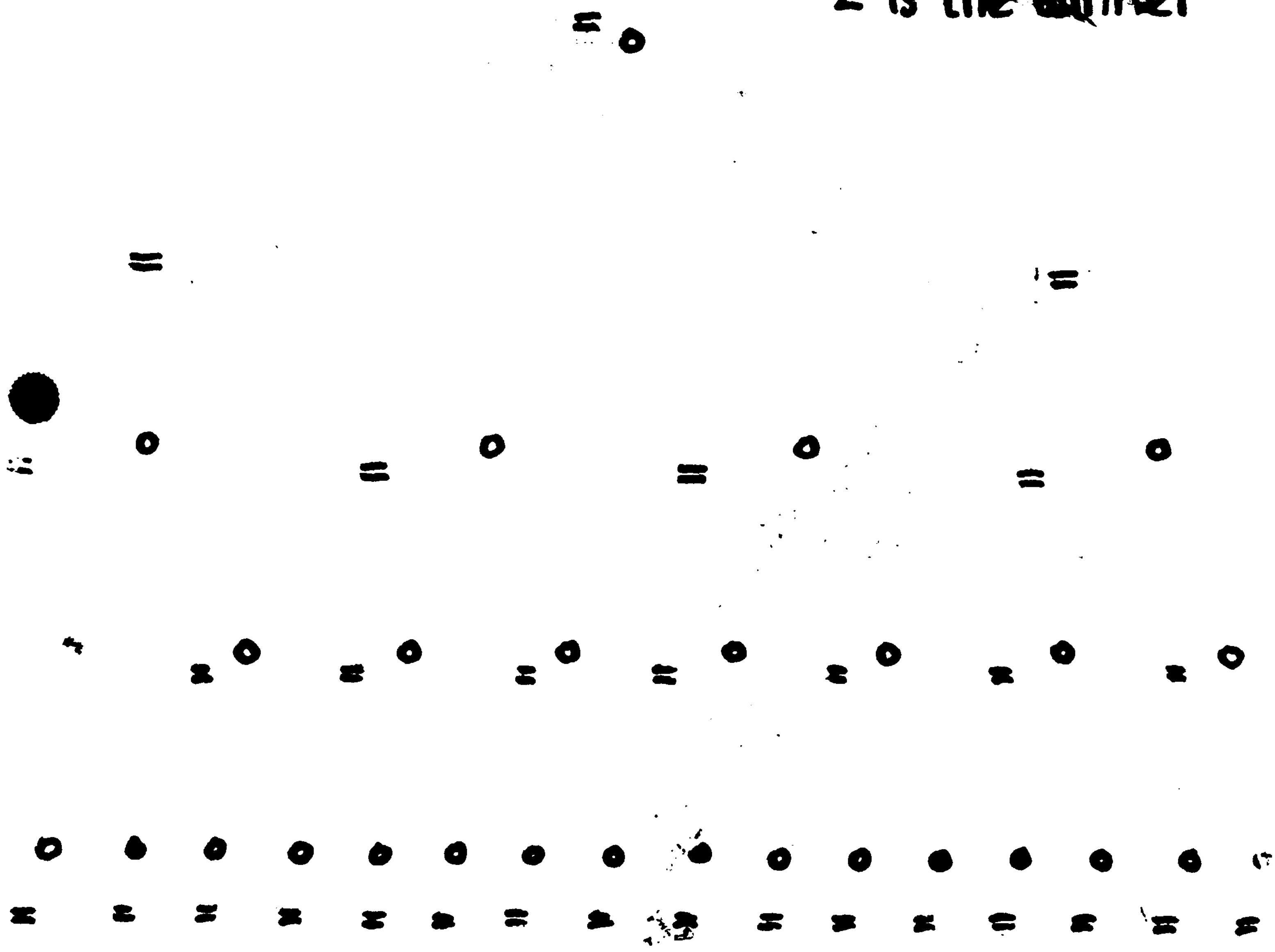
SOPHISTICATED VOTING (UNDER AMENDMENT PROCEDURE.

Definition. The anticipated decision at a vertex in level $m-1$ is the alternative labelling that vertex. For $0 \leq j < m-1$, the anticipated decision at a vertex in level j is the majority choice (given by T) between the 2 alternatives which are the anticipated decisions at the 2 vertices at level $j+1$ which are dominated by v.

The sophisticated decision is the anticipated decision at the root (level 0).

Sophisticated majority voting with agenda 524316

2 is the winner



PROPOSITION:

Let T be a majority tournament on $\{a_1, a_2, \ldots, a_m\}$, let $A = (a_1, a_2, \ldots, a_m)$ be an agenda, and let x denote the sophisticated decision. Then

i) if $x \neq a_m$, then $x \rightarrow a_m$,

ii) if $y \in T$ is a transmitter, then $y = x$, and

iii) $x = a_m$ if and only if a_m is a transmitter.

The Shepsle-Weingast Algorithm :

Given a majority tournament T on $\{a_1, \ldots, a_m\}$ and an agenda $A = (a_1, \ldots, a_m)$ form the sophisticated sequence $Z = (z_1, \ldots, z_m)$ as follows:

$$z_m = a_m$$

$$1 \leq i < m : \quad z_i = \begin{cases} a_i, & \text{if } a_i \to z_{i+1}, z_{i+2}, \ldots, z_m \\ z_{i+1}, & \text{otherwise} \end{cases}$$

THEOREM (Shelpsle-Weingast, 1984):
z_1 is the sophisticated decision.

Example:

```
      5 2 4 3 1 6
                6
              1 6
            1 1 6
        4   1 1 6
      2 4   1 1 6
    2 2 4   1 1 6
    ↑
```

Proposition: Let T be a majority tournament on $\{a_1, \ldots, a_m\}$ and let $A = (a_1, \ldots, a_m)$ be an agenda. If

a_i = the sophisticated decision, and

a_j = the sincere decision,

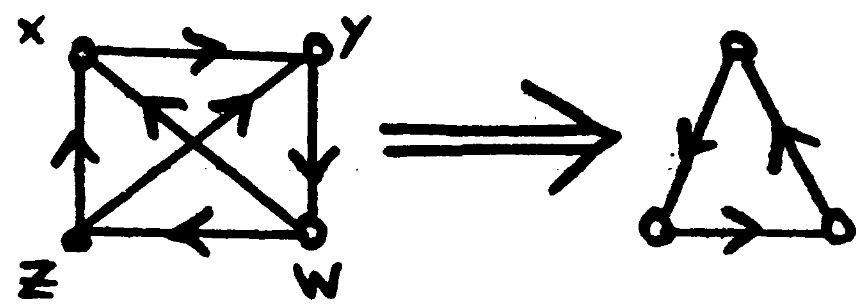
then either

$i = j$, or

$i < j$ and $a_i \rightarrow a_j$ in T.

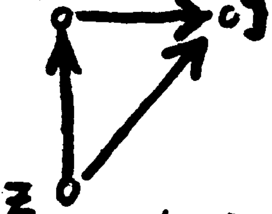
Where in T are the possible sophisticated decisions?

A Banks point is a vertex x in T which is a transmitter of some transitive subtournament W of T such that no vertex of T dominates all vertices of W.



Banks points = {z, w, y}

For example: Use W :



This example shows Banks points ≠ T

- THEOREM (Banks, 1985)

  Alternative a is a sophisticated decision if and only if a is a Banks point.

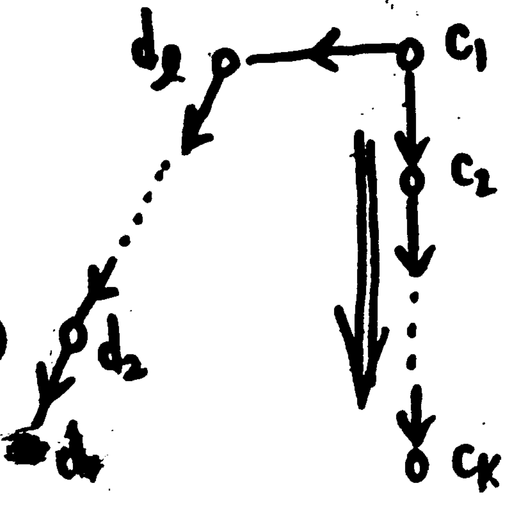
<u>Example</u>: T a tournament with a transmitter x:

i) $T^* = \{x\}$, so x is the sincere dec.

ii) x is the only Banks point, so x is the sophisticated dec.

<u>Another Example</u>: Pick a maximal transitive subtournament in tournament T, say given by $c_1 \to c_2 \to \cdots \to c_k$. Let $d_\ell \to d_{\ell-1} \to \cdots \to d_1$ be a spanning path in $T - \{c_1, c_2, \ldots, c_k\}$.

ASSUME $c_1 \to d_\ell$



Use agenda

$(d_1, d_2, \ldots, d_\ell, c_1, c_2, \ldots, c_k)$

c_1 is the sincere dec.

c_i is the sophisticated dec.

Of course, if T^* has an agenda for which the sincere dec. and the sophisticated dec. are the same then so does T.

However,  has no such agenda!

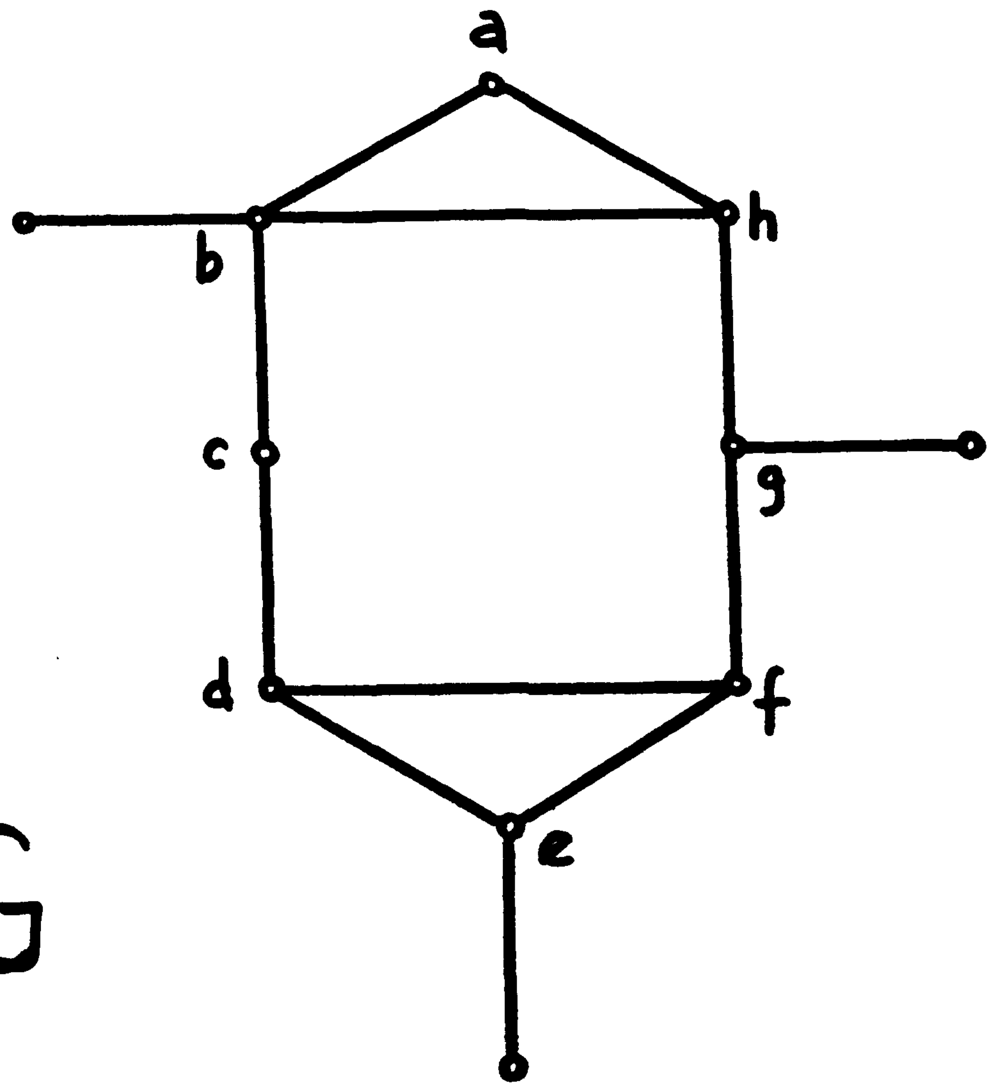
And, any T with $T^* \equiv$  has no such agenda.

Problem: For which tournaments is there an agenda containing an alternative which is both the sincere and the sophisticated decision?

K. B. Reid — Majority Tournaments: Sincere and sophisticated voting decisions under amendment procedure, Math. Soc. Sci. 21 (1991) 1–19.

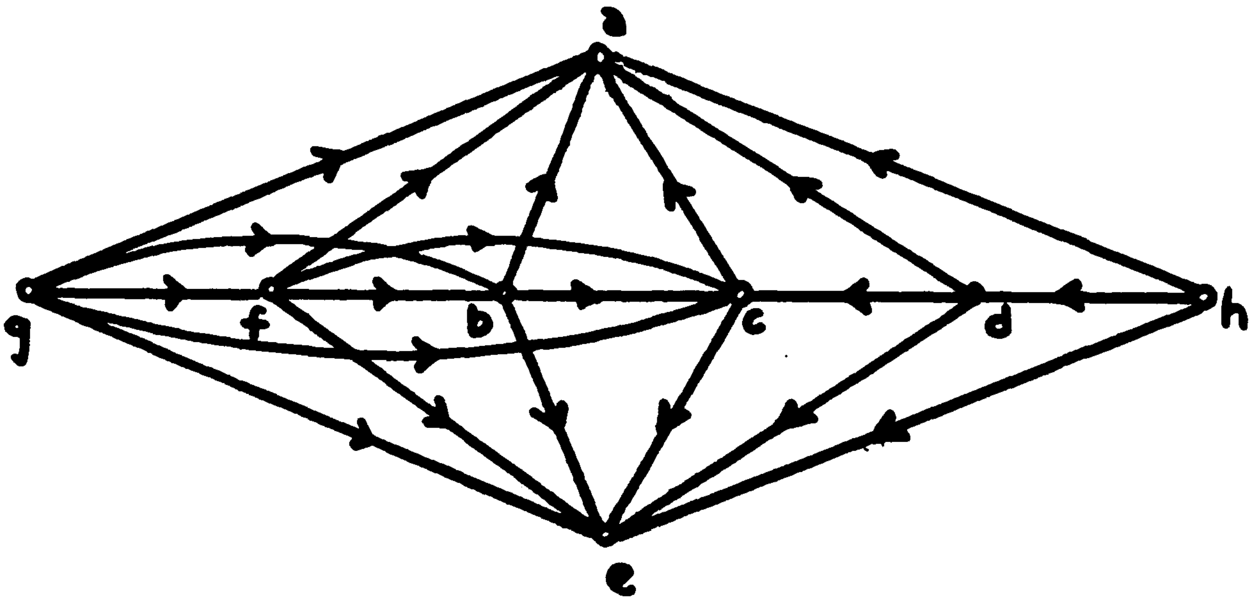
K. B. Reid — The relationship between two algorithms for decisions via sophisticated majority voting with an agenda, Disc. Appl. Math. 31 (1991) 23-2

ORIENTED GRAPHS REALIZED BY DISTANCE-PLURALITIES IN TREES



G

~~The~~ purple vertices are candidate locations. They form the vertex set of a digraph in which (x,y) is an arc iff more vertices of G are closer to x than to y.



This digraph is called the plurality preference digraph of G with respect to V (the set of vertices of G, in this example) and C (the set of candidate locations).

This digraph is also said to be the realization of G with respect to V and C.

Formal Definition

Let G be a connected graph of order n. Let V and C denote two (not necessarily disjoint) subsets of vertices of G, $|V| = v$ and $|C| = c$.

The realization of G with respect to V and C is the digraph with vertex set C so that (a, b) is an arc iff

$$|\{x : x \in V \text{ and } d_G(x,a) < d_G(x,b)\}|$$
$$> |\{x : x \in V \text{ and } d_G(x,b) < d_G(x,a)\}|.$$

Suppose that we are given an asymmetric digraph D, say of order c.

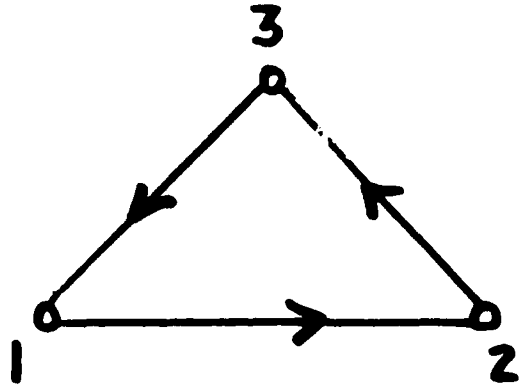
- Is there a connected graph G of order n which contains two subsets V and C, $|V| = v$ and $|C| = c$, so that D is the [illegible] of G w.r.t. V and C?

Yes. (T. [illegible] and [illegible])

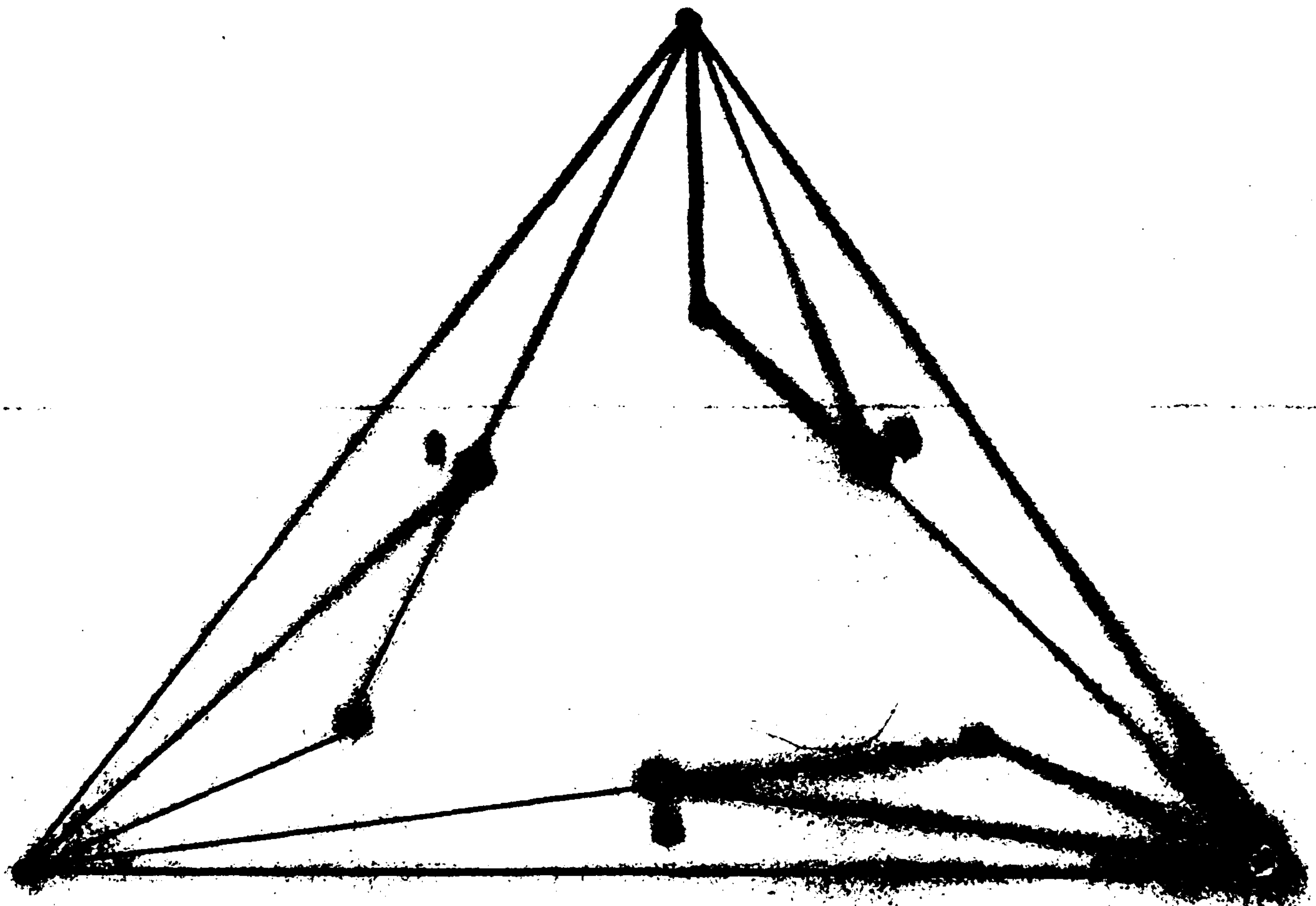
Def. [illegible]

[illegible]

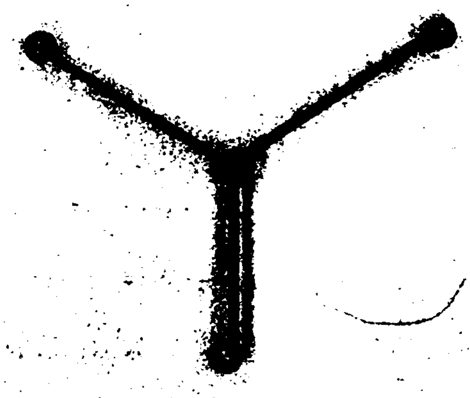
[illegible]



Γ



D is (1,0,3)-[illegible]



[redacted] realiza [redacted]
([redacted],10,5)

Let D be an oriented graph of order c with q arcs and maximum degree Δ.

- D is (n, n, c)-realizable, where $n = c^2 + c\Delta - q$. (Johnson & Slater. [illegible]

- D is (n, n, c)-realizable, where $n = \min(c^2+2, 2c+2q)$.

  ([illegible])

- D is $(3c, 2c, c)$-[illegible].

  (W. Schnyder)

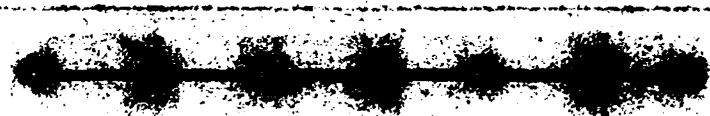
- D is ([illegible], [illegible], c)-r[illegible].

Which oriented graphs are realizable by TREES?

Theorem. If an oriented graph D is (n,n,c)-realizable by a TREE, then D is transitive.

Example: The oriented graph of order 8 above is not $(n,n,8)$-realizable by a tree for all $n \geq 8$.

Theorem. If an oriented graph D is (m.n.c)-realizable by a TREE, then D contains no induced anti-directed path of length 3.



IN FACT.

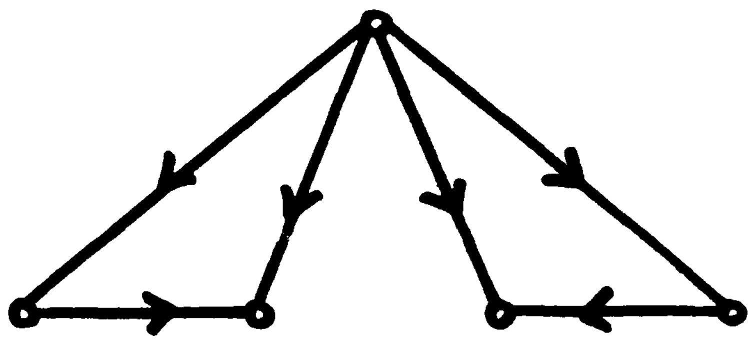
THEOREM. Let D be an oriented graph of order c.

D is (n, n, c)-realizable by a TREE of order $n > c$

if and only if

D is transitive and D contains no induced anti-directed path of length 3.

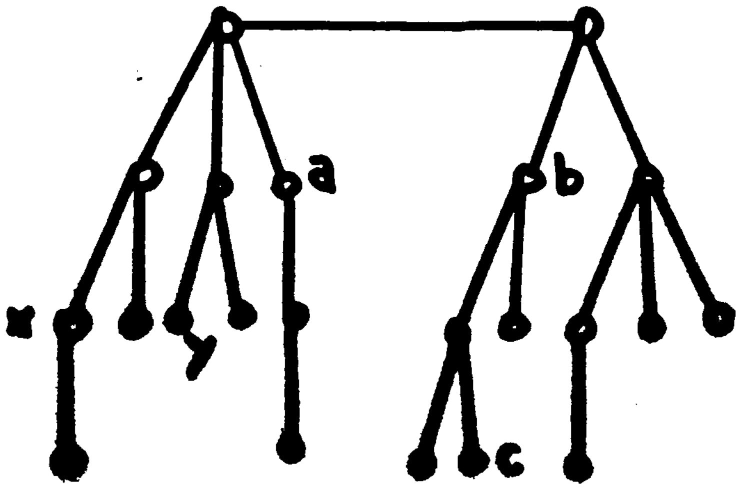
REMARK:



is transitive & contains no a-d path of length 3

BUT

is NOT realizable by any TREE of order 5!!



$\exists$ an arc from a to c in D

$\exists$ no arc between a and b in D

Let k be a non-negative integer, and let c be a vertex in a graph G

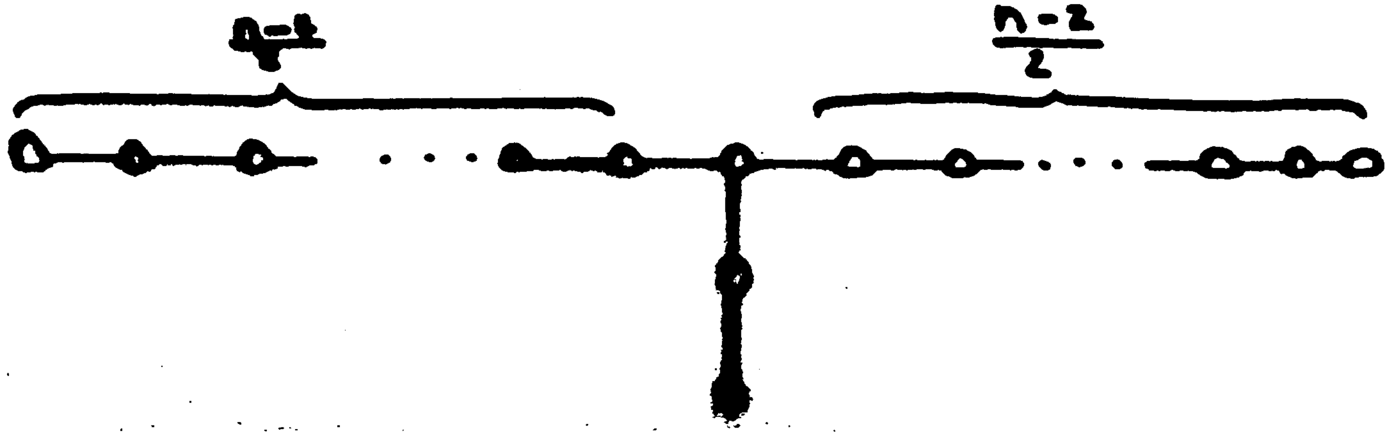
$$B_G(c, k)$$

denotes the ball of radius k about c.

i.e. $B_G(c, k) = \{ x \in V(G) \; ; \; d_G(c, x) \leq k \}$.

<u>THEOREM</u> Let T be a tree of order n with a single centroid vertex c. If for each k, $0 \leq k < \mathrm{dia}(T)$, $T - B_T(c, k)$ consists of subtrees with distinct orders, then the transitive n-tournament is (n, n, n)-realizable by T.

$n \geq 10$, n even :



$n \geq 7$, n odd :

[illegible]

CENTRALITY IN GRAPHS

- TREES
- "FRINGE" VERTICES

STANDARD MEASURES OF CENTRALITY IN TREES:

CENTER (C. JORDAN, 1869)

MEDIAN (O. ORE, 1962)

CENTROID (JORDAN)

branch weight centroid

$$e(x) = \max_{v \in V(T)} d(x,v)$$

center of $T = \{x \,;\ e(x) \le e(u) \text{ for all } u \in V(T)\}$

$$\Delta(x) = \sum_{v \in V(T)} d(x,v)$$

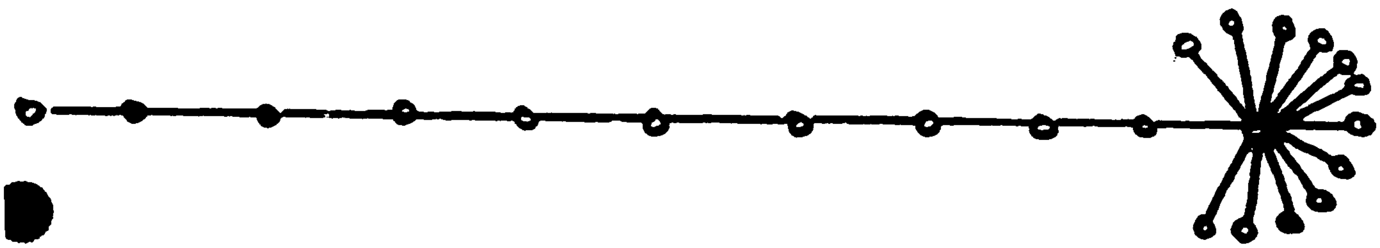
median of $T = \{x \,;\ \Delta(x) \le \Delta(u) \text{ for all } u \in V(T)\}$

$bw(x) =$ order of a largest component of $T - x$

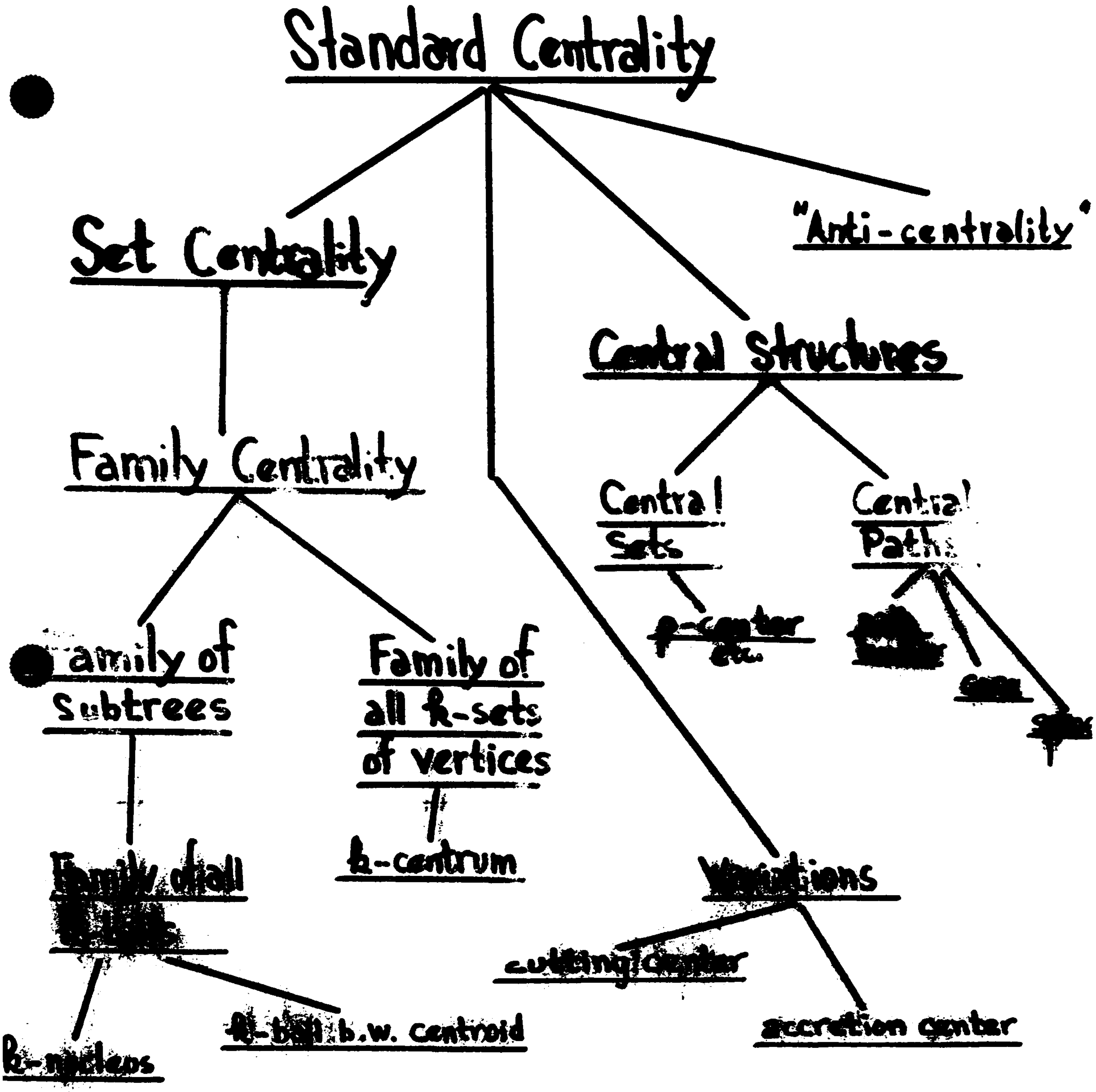
centroid of $T = \{x \,;\ bw(x) \le bw(u) \text{ for all } u \in V(T)\}$

— FUNDAMENTAL RESULTS FOR TREES —

Theorem 1. (Jordan, 1869) The center and branch-weight centroid each consists of a single vertex or two adjacent vertices.



Theorem 2. (Zelinka, 1968) The median is the same as the branch-weigh[illegible] centroid.



Standard Centrality
Set Centrality
"Anti-centrality"
Central Structures
Family Centrality
Central Sets
Central Paths
Family of Subtrees
Family of all k-sets of vertices
k-centrum
b.w. centroid
accretion center

Family-centrality (Slater 1981)

- Let $\mathcal{S} = \{S_1, \ldots, S_m\}$ be a family of connected subgraphs of connected graph G which has n vertices.

$$E_{\mathcal{S}}(x) \equiv \max_{1 \le i \le m} d(x, S_i) \qquad \min_{w \in S_i} d(x,w)$$

$\mathcal{S}$-center of $G \equiv \{x \,;\, E_{\mathcal{S}}(x) \le E_{\mathcal{S}}(y)$ for all $y \in V(G)\}$

- $s_{\mathcal{S}}(x) \equiv \sum \{ d(x, S_i) \,;\, 1 \le i \le m \}$

$\mathcal{S}$-median of $G \equiv \{x \,;\, s_{\mathcal{S}}(x) \le s_{\mathcal{S}}(y)$ for all $y \in V(G)\}$

- $w_{\mathcal{S}}(x) \equiv$ max. no. of mem[illegible]
contained in [illegible] $G-x$

[illegible]

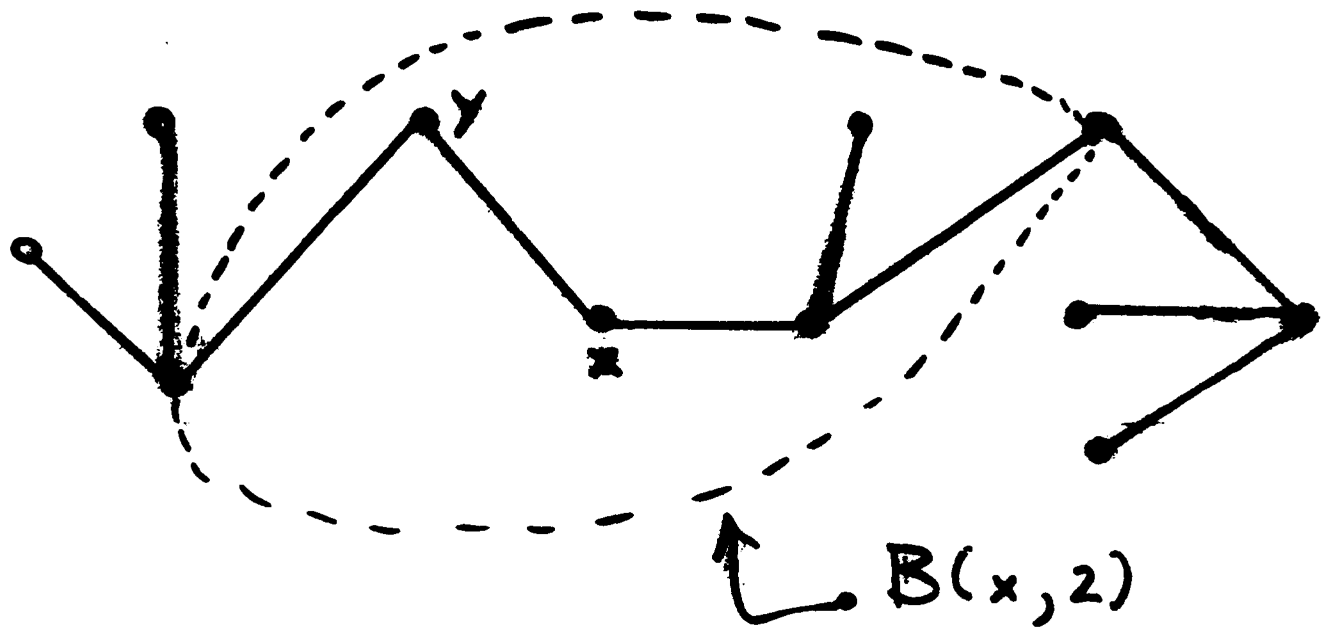
Assume: G a tree T and $\mathcal{S}$ a collection of subtrees with no common vertex

Theorem 3′. The $\mathcal{S}$-center of T consists of one vertex or two adjacent vertices.

Theorem 4′. The $\mathcal{S}$-median of T consists of the vertices of a path in T.

Theorem 5′. The $\mathcal{S}$-median of T is contained in the $\mathcal{S}$-branch weight centroid of T (Containment can be proper.)

For $x \in V(T)$ and for integer $k \geq 0$, the k-ball about x, denoted $B(x,k)$ is $\{z \,;\, z \in V(T),\ d(x,z) \leq k\}$.



$\sum \{ d(u, B(x,2)) : u \in V(T) \} = 7$

$\sum \{ d(u, B(y,2)) : u \in V(T) \} = 10$

- Observe that if $\mathcal{S}$ is the family of all k-balls* in T, $k \le r(T)$, then the $\mathcal{S}$-center of T is the same as the center of T.

- Because, for $r(T) \ge k$,

$$
\begin{aligned}
E_{\mathcal{S}}(x) &= \max_{u \in V} d(x, B(u,k)) \\
&= \max_{u \in V} d(u, B(x,k)) \\
&= \max_{u \in V} (d(u,x) - k) \\
&= e(x) - k
\end{aligned}
$$

* i.e. subgraphs induced by all the k-balls

— RELAX CONNECTIVITY REQUIREMENT ON FAMILY MEMBER

Let $\mathcal{S} = \{S_1, \ldots, S_m\}$, $m = \binom{n}{k}$,

be the family of all k-subsets of

the the n-set of vertices of a tree T.

$E_{\mathcal{S}}(x) \equiv \max\{d(x,S_1), \ldots, d(x,S_m)\}$

What is the $\mathcal{S}$-center in this case?

$bw_{\mathcal{S}}(x) \equiv$ max. no. of members of $\mathcal{S}$ which are entirely contained in one component of $T - x$.

$\mathcal{S}$-branch weight centroid $\equiv$ ordinary b-w centroi

For $x \in V$ and k a non-negative integer

let $\rho(x, k)$ denote the sum

$\sum \{ d(u, B(x,k)) : u \in V \}$, where

$d(u, B(x,k)) = \min \{ d(u,z) : z \in B(x,k) \}$.

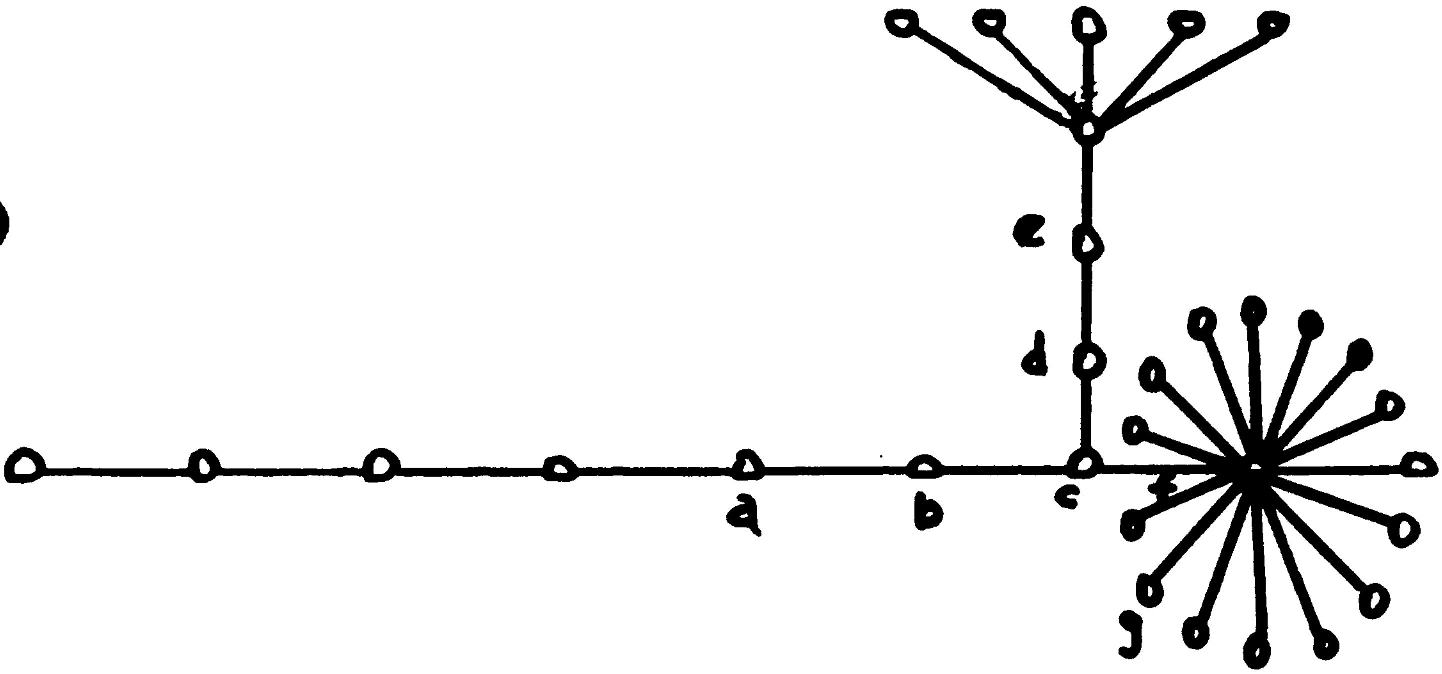
The k-nucleus of T is the subset of vertices

$\{ x : x \in V, \rho(x,k) \leq \rho(y,k) \ \forall y \in V \}$;

it is denoted $\Theta(T; k)$.

RE: – $\Theta(T; 0)$ is the <u>median of T</u>

– if $r = \text{radius}(T)$, then $\Theta(T; r)$ is the <u>center of T</u>



~ $p(\cdot, k)$ ~

a	b	c	d	f	e	g	$\Theta(T; k)$
118	97	78	93	77	110	106	f
88	67	48	63	47	80	76	f
60	39	21	35	33	52	47	c
34	14	11	10	21	30	33	d
11	6	3	6	11	10	21	c
5	0	1	3	3	6	11	b
0	0	0	1	1	3	3	a, b, c

EXAMPLE DUE TO P. SLATER (198

Observe that since

$$d(u,x) = d(x,u),$$

$$d(u, B(x,k)) = d(x, B(u,k)).$$

So,

$$p(x,k) = \sum \{ d(x, B(u,k)) \, ; \, u \in V \}$$

$$= s_{\mathcal{S}}(x),$$

where $\mathcal{S} = \{ B(u,k) \, ; \, u \in V \}$.

Thus, the k-nucleus of T is the $\mathcal{S}$-median of T, where $\mathcal{S}$ is the family of all k-balls in T.

Peter J. Slater (1981) showed:

1. For $0 \leq k \leq \text{rad}(T)$, $\Theta(T;k)$ consists of a single vertex or two adjacent vertices.

2. $\bigcup \{\Theta(T;k) : 0 \leq k \leq \text{rad}(T)\}$ induces a subtree of T.

F. For $x \in V$ and $S \subseteq V$, $S \neq \emptyset$, let

$\sigma(x;S) = \sum_{s \in S} d(x,s)$. For positive integer k define $r(x;k)$ to be the number

$\max\{\sigma(x;S) : S \subseteq V, |S| = k\}$.

The k-centrum of T, denoted $C(T;k)$ consists of all x for which $r(x;k)$ is a minimum, i.e.

$$C(T;k) = \{x : x \in V, r(x;k) \leq r(y;k) \ \forall y \in V\}$$

NOTE: – $C(T;1)$ is the center of T

– $C(T;n)$ is the median (= bw centroid) of T, where $n = |V|$.

~ $r(\cdot, k)$ ~

k	a	b	c	d	f	e	g	$C(T;k)$
1	6	5	6	7	7	8	8	b
2	12	10	11	13	13	15	15	b
3	18	15	15	18	18	21	21	b, c
4	24	20	19	22	23	26	27	c
5	30	25	23	25	28	30	33	c
6	35	30	27	28	33	34	39	c
7	39	34	31	31	38	38	45	c, d
8	43	38	35	34	43	42	51	d
9	47	41	38	37	47	46	56	d
10	51	44	41	40	51	50	61	d
11	55	47	43	43	54	54	65	c, d
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
24	106	86	69	82	71	100	95	c
25	109	89	71	85	72	102	97	c
26	112	91	73	87	73	104	99	c, f
27	114	93	75	89	74	106	101	f
28	116	95	76	91	75	108	103	f
29	117	96	77	92	76	109	105	f
30	118	97	78	93	77	110	106	f
31	118	97	78	93	77	110	106	f

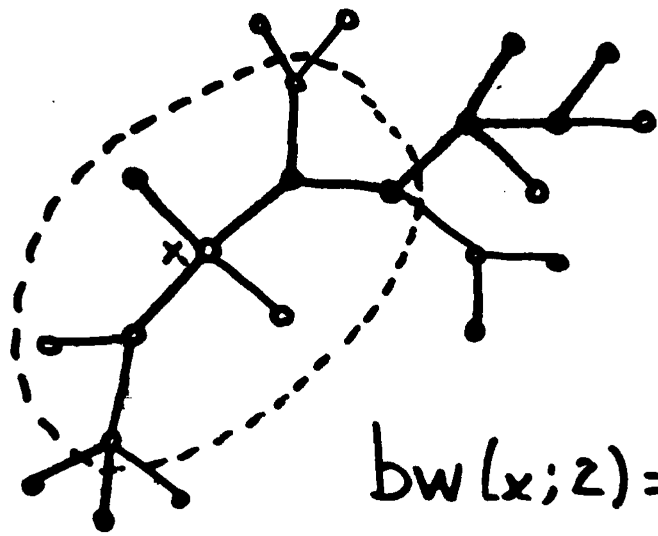
- Peter J. Slater (1978) showed that:

1. For $1 \leq k \leq |V|$, $C(T:k)$ consists of a single vertex or 2 adjacent vertices.

2. $\bigsqcup_{1}^{|V|} C(T;k)$ induces a sub<u>tree</u> of T.

LET $T=(V,E)$ BE A TREE, $|V|\geq 2$:

DEF. For $x\in V$ and non-negative integer k, let $B(x,k)$ denote the k-ball about x, i.e. $\{z : z\in V, d(x,z)\leq k\}$, and let $bw(x;k)$ denote the numbe of vertices in a largest component of $T-B(x,k)$.

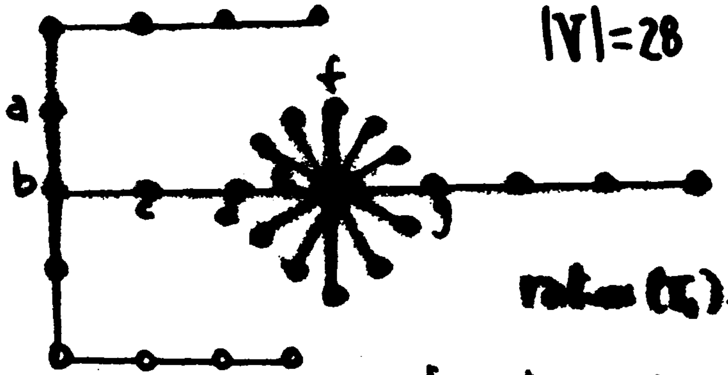


$bw(x;2)=6$

$bw(x;k)$ is called the k-ball branch weight of x.

DEF. The k-ball branch weight centroid of T, denoted $W(T;k)$, consists of all vertices of T of minimum k-ball branch weight.

EXAMPLE 1.



$|V| = 28$

T_1

$\mathrm{rad}(T_1) = 6$

	a	b	c	d	e	f	g	$W(T_1;k)$
$bw(\cdot;0)$	23	17	16	15	13	27	24	e
$bw(\cdot;1)$	17	16	15	14	12	13	13	d
$bw(\cdot;2)$	16	15	4	5	[illegible]	12	12	c
$bw(\cdot;3)$	15	4	3	4	[illegible]	[illegible]	11	c
$bw(\cdot;4)$	4	3	2	3	4	5	5	c
$bw(\cdot;5)$	3	2	1	2	3	4	4	c
$bw(\cdot;6)$	2	1	0	1	2	3	3	c
$bw(\cdot;7)$	1	0	0	0	1	2	2	b, c, d

NOTE:

$W(T;0)$ is the usual branch weight centroid

$W(T;r)$ is the usual center. where $r = \mathrm{rad}($

For $k \geq \mathrm{rad}(T)$, $W(T;k+1) = W(T;k) \cup \left(\bigcup_{x \in W(T;k)} B(x,1) \right)$

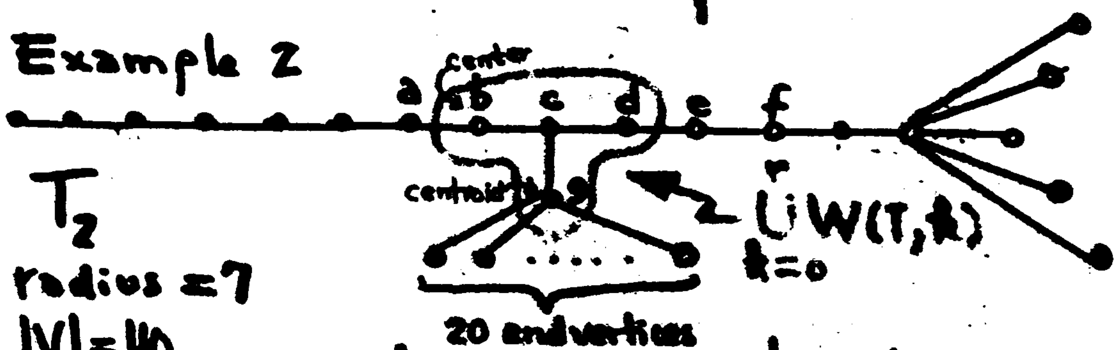
So, for $k \geq \mathrm{dia}(T)$, $W(T;k) = V$.

THEOREM. For $0 \leq k \leq \mathrm{rad}(T)$, $W(T;k)$ consists of a single vertex or two adjacent vertices.

THEOREM. $\bigcup\{W(T:k) \mid 0 \leq k \leq rad(T)\}$ induces a subtree of T.

Observation: $\bigcup\{W(T:k) \mid 0 \leq k \leq rad(T)\}$ need not induce a path in T !

Example 2



T_2
radius $= 7$
$|V| = 40$

	a	b	c	d	e	f	g	$W(T_2:k)$
bw(·,0)	33	32	21	30	31	32	19	g
bw(·,1)	32	21	9	21	30	31	10	c
bw(·,2)	21	9	8	7	21	30	9	d
bw(·,3)	9	8	7	6	7	21	8	d
bw(·,4)	8	7	6	5	6	7	7	d
bw(·,5)	7	6	3	4	5	6	6	c
bw(·,6)	6	1	2	3	4	5	3	b
bw(·,7)	1	0	1	2	3	4	2	b

B. Zelinka (1967) showed :

the branch weight centroid of T =

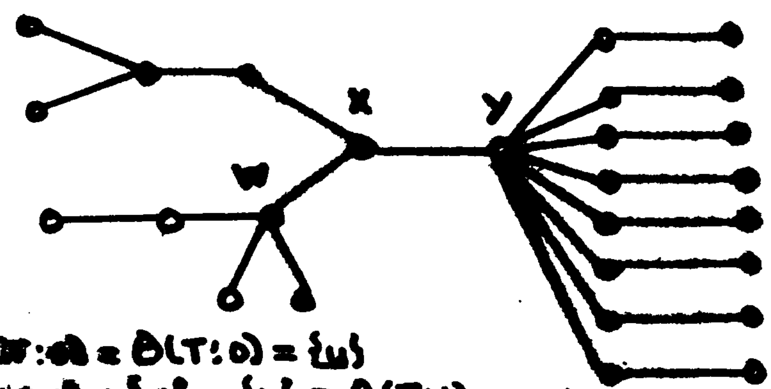
the median of T

That is,

$$W(T;0) = \theta(T;0).$$

Also, $W(T;r) = \theta(T;r) =$ center.

Must $W(T;k) = \theta(T;k)$ for $0 < k < \mathrm{rad}(T)$?



radius = 3

$W(T;0) = \theta(T:0) = \{w\}$
$W(T;1) = \{x\}$ [illegible] $\{y\} = \theta(T;1)$
$W(T;2) =$ [illegible] $\{y\} = \theta(T;2)$
$W(T;3) = \theta(T;3) = \{x\}$

i.e. $\left(\bigcup_{1}^{r-1} W(T:k)\right) \cap$ [illegible] is empty for this T.

SOME "ANTI-CENTRALITY"

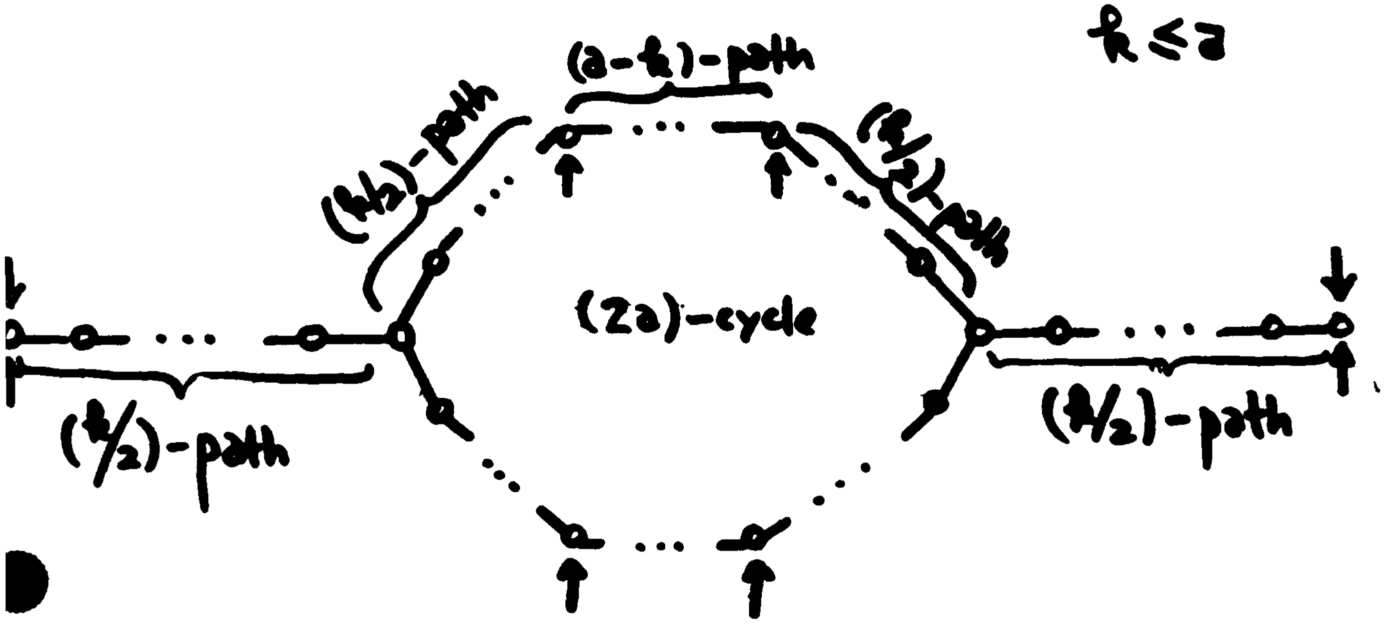
Weizhen Gu

$P(G) \equiv \{x : x \in V(G),\ e(x) = \mathrm{dia}(G)\}$

$EC(G) \equiv \{z : z \in V(G),\ d(z,c) = r(G)$ for some center vertex $c\}$

WHEN IS $P(G) = EC(G)$?

Certainly $P(T) = EC(T)$, for trees T.



In fact, for any two positive integers $p<q$, there exists a G with

$$\frac{d(P(G), EC(G))}{dia(G)} = \frac{p}{q}.$$

The tree result above has been extended in two directions (Gu and Reid, 1992)

If the center of G is a single vertex v, and if any block containing v is complete, then $P(G) = EC(G)$.

G has at least 2 center vertice

If G is not self-centered, and the subgraph induced by the center of G is contained in a block that is complete, $P(G)=EC(G)$.

- The other direction concerns some nesessary and sufficient conditions on some graphs G with $\mathrm{dia}(G) = 2r(G)$ or $2r(G)-1$ so that $P(G) = EC(G)$.

K.B. Reid, Centroids to centers in trees, NETWORKS 21 (1991) 11–17.

K.B. Reid and Weizhen Gu:

- Plurality preference digraphs realized by trees. I: Necessary and sufficient conditions, submitted.
- Plurality preference digraphs realized by trees. II: On realization numbers, to appear in Discrete Mathematics (1992).
- Peripheral and eccentric vertices in graphs, to appear in Graphs and Combinatorics (1992)
- (with W. Schnyder) Realization of digraphs by preferences based on distances in graphs, submitted.

CLUSTER ANALYSIS ALGORITHMS

Pierre Hansen
GERAD and École des Hautes Études Commerciales
Département des Méthodes Quantitatives et Systèmes d'Information
5255 avenue Decelles
Montréal, Canada H3T 1V6

Brigitte Jaumard
GERAD and École Polytechnique de Montréal
Département de Mathématiques Appliquées
Succursale A, Case Postale 6079
Montréal, Canada H3C 3A7

Cluster analysis addresses the following very general problem: given a set of entities. find subsets of it, called *clusters*, which are *homogeneous*. i.e., such that entities within the same cluster resemble one another, and/or *well separated*. i.e.. such that entities in different clusters differ one from the other.

From the late 60's, Mathematical Programming has been applied to cluster analysis. This allowed to:

(i) formulate precisely many problems of cluster analysis as mathematical programs: i.e., optimization problems with an explicit objective function and constraints;

(ii) study the computational complexity of these problems;

(iii) obtain new polynomial algorithms for easy problems, with a low, and sometimes lowest possible, complexity by careful study of each of their steps and of the data structures necessary to their implementation;

(iv) obtain new and practically efficient algorithms for NP-hard problems;

(v) derive theoretical properties of existing or new algorithms;

(vi) make new clustering methods available to researchers in various fields in the form of computer packages.

Our work in the last few years and planned for follows and extends such lines. We explore two avenues of research. On one hand, we define several new problems of cluster analysis in the areas of divisive hierarchical clustering, partitioning, constrained partitioning and clustering with asymmetric dissimilarities. Each time, complexity issues are or will be explored and new exact algorithms designed (as well as heuristics in the cases where exact solution is too time-consuming). The main tools used will be graph theory, combinatorial optimization (mainly nonlinear 0–1 programming) as well as, for the first time in cluster analysis, exact methods of global optimization.

On the other hand, we complement the study of algorithms by analysis of the steps preceding and following their use: ways to construct dissimilarity indices, automated methods to select "best" partitions among efficient ones, derivation of robustness and sensitivity concepts for clusters and partitions, new and more informative ways to represent partitions and hierarchies. This work will be done in parallel with the analysis of several data sets for real world problems.

The first part of the presentation will review steps of a cluster analysis study and illustrate the mathematical programming approach. Bicriterion cluster analysis with the split and diameter criteria will be used for this purpose. The second part will review recent work on average linkage divisive hierarchical clustering, maximum split clustering with connectivity constraints, the Weber problem on the sphere and using espaliers to represent results.

PARTITIONING PROBLEMS AND ALGORITHMS IN CLUSTER ANALYSIS

PIERRE HANSEN, RUTCOR, RUTGERS UNIVERSITY AND GERAD, ECOLE DES HAUTES ETUDES COMMERCIALES, MONTREAL.

"... le seul moyen de faire une methode instructive et naturelle, c'est de mettre ensemble les choses qui se ressemblent et de séparer celles qui diffèrent les unes des autres."

BUFFON, "HISTOIRE NATURELLE",
PREMIER DISCOURS.

CLUSTERS (OR CLASSES) SHOULD BE

- HOMOGENEOUS
- WELL SEPARATED

I. CLUSTER ANALYSIS: DEFINITION, STEPS OF A STUDY, PROBLEMS

DATA ANALYSIS : TECHNIQUES FOR ANALYSING MULTIDIMENSIONAL DATA

- PRINCIPAL COMPONENTS ANALYSIS
- DISCRIMINANT ANALYSIS
- CLUSTER ANALYSIS
- ....

BASIC PROBLEM OF CLUSTER ANALYSIS :

GIVEN A SET OF ENTITIES, FIND SUBSETS, CALLED CLUSTERS, WHICH ARE HOMOGENEOUS AND/OR WELL SEPARATED

HOMOGENEITY : ENTITIES IN THE SAME CLUSTER SHOULD RESEMBLE THEMSELVES

SEPARATION : ENTITIES IN DIFFERENT CLUSTERS SHOULD NOT RESEMBLE THEMSELVES (BUFFON 1759, HISTOIRE NATURELLE)

THESE CONCEPTS CAN [illegible] MATHEMATICALLY, IN MA[illegible]

USES OF CLUSTER ANALYSIS

- CLASSIFICATION (PLANTS, ANIMALS)
- EXPLORATION : FIND A STRUCTURE IF THERE IS ONE (NATURAL AND SOCIAL SCIENCES)
- PREDICTION : BEHAVIOUR OF ENTITIES WITHIN CLUSTERS (MEDECINE : DIAGNOSIS, MARKETING)
- ORGANIZATION : OPERATIONAL AIM (OPERATIONS RESEARCH, APPLIED ECONOMICS, PRODUCTION ...)

EXAMPLE 1 : STANDARD SIZES FOR CLOTHING

EXAMPLE 2 : POLITICAL DISTRICTING

EXAMPLE 3 : GROUPING PART TECHNOLOGY

STEPS OF A CLUSTER ANALYSIS STUDY

(i) SAMPLE $O = \{ o_1 , o_2 , \dots \; o_N \}$

(ii) OBSERVATIONS OR MEASUREMENTS

$$X \equiv (x_{it}) \qquad \begin{array}{l} i = 1, \dots, N \\ t = 1, \dots, q \end{array}$$

$\Rightarrow$ RAW DATA

(iii) DISSIMILARITIES

$$D \equiv (d_{k\ell}) \qquad \begin{array}{l} k = 1, \dots N \\ \ell = 1 \dots N \end{array}$$

$d_{k\ell} \geq 0$, $d_{kk} = 0$, $d_{\ell k} = d_{k\ell}$ (USUALLY)

(NOT NECESSARILY A DISTANCE :

$d_{kp} \leq d_{k\ell} + d_{\ell p}$ MAY NOT HOLD)

THE MORE o_k AND o_ℓ ARE DIFFERENT, THE LARGER IS $d_{k\ell}$.

MANY FORMULAE AVAILABLE (e.g. EUCLIDEAN DISTANCE , JACCARD DISTANCE ...) SOME AXIOMATIC STUDIES (e.g. BEAULIEU, [illegible])

(iv) STRUCTURE FOR THE CLUSTERS

(iv, i) SUBSET $C \subset O$

(iv, ii) PARTITION P_M OF O INTO M CLUSTERS

$P_n = \{ c_1, c_2, \ldots\ c_n \}$

$c_i \neq 0 \quad , \quad c_i \cap c_j = \phi \quad , \quad \bigcup_{j=1..n} c_j = 0$

(iv, iii) HIERARCHY OF PARTITIONS

$P_N , P_{N-1} , \ldots\ P_1$

WITH $c_j \in P_k$, $c_\ell \in P_r$ AND $r < k$

$\Rightarrow c_j \cap c_\ell = \phi$ OR $c_j \subset c_\ell$

A HIERARCHY OF PARTITIONS IS DEFINED BY 2N-1 CLUSTERS WHICH ARE PAIRWISE DISJOINT OR INCLUDED ONE INTO THE OTHER.

LESS FREQUENTLY : COVERING , HIERARCHY OF COVERINGS

(POSSIBLY ADDITIONAL CONSTRAINTS : CONNECTIVITY , CARDINALITY , WEIGHT , BALANCING

(V) CRITERION

MANY POSSIBLE CRITERIA (DISCUSSED BELOW) THEY MAY BE GROUPED AS :

- THRESHOLD ~~TYPE~~ : VALUE GIVEN BY A SINGLE DISSIMILARITY

EXAMPLE : DIAMETER OF A PARTITION =
LARGEST DISSIMILARITY BETWEEN TWO ENTITIES
IN THE SAME CLUSTER

- THRESHOLD SUM TYPE : VALUE GIVEN BY ONE DISSIMILARITY FOR EACH CLUSTER

EXAMPLE : SUM OF DIAMETERS OF A PARTITION

- SUM TYPE : VALUE GIVEN BY DISSIMILARITIES BETWEEN ENTITIES IN ONE CLUSTER

EXAMPLE : $\underbrace{\text{SUM}}^{\text{LARGEST}}$ OF DISSIMILARITIES IN ONE CLUSTER (P-MEDIAN MAX PROBLEM)

- SUM-SUM TYPE : VALUE DEPENDS ON ALL DISSIMILARITIES

EXAMPLE : MINIMUM WITHIN CLUSTERS VARIANCE

$\Rightarrow$ WELL-POSED MATHEMATICAL PROGRAM

(vi) ALGORITHM

EXACT OR HEURISTIC, POLYNOMIAL OR NOT

(vii) INTERPRETATION OF RESULTS

EXISTENCE OF A "NATURAL" STRUCTURE, DESCRIPTION OF CLUSTERS.

2. THRESHOLD TYPE CRITERIA

- SPLIT OF A CLUSTER : SMALLEST DISSIMILARITY BETWEEN AN ENTITY INSIDE IT AND ONE OUTSIDE

$$s(C_j) = \min_{k \mid O_k \in C_j} \; \min_{l \mid O_l \notin C_j} d_{kl}$$

DIAMETER OF A CLUSTER : LARGEST DISSIMILARITY BETWEEN TWO ENTITIES IN THE CLUSTER

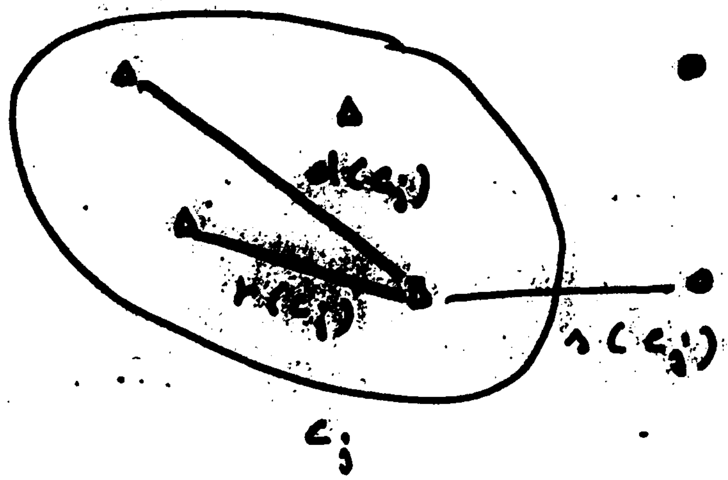
$$d(C_j) = \max_{k,l \mid O_k, O_l \in C_j} d_{kl}$$

RADIUS OF A CLUSTER : MINIMUM OVER ALL ENTITIES OF THE CLUSTER OF THE MAXIMUM DISSIMILARITY BETWEEN THAT ENTITY AND ANOTHER ONE IN THE CLUSTER

$$r(C_j) = \min_{k \mid O_k \in C_j} \; \max_{l \mid O_l \in C_j} d_{kl}$$

(O_k ≡ CENTER OF C_j

EX : POINTS IN THE PLANE ; EUCLIDEAN DISTANCE



SPLIT OF A PARTITION P_M : MINIMUM SPLIT OF THIS PARTITION'S CLUSTERS (OR MINIMUM DISSIMILARITY BETWEEN ENTITIES IN DIFFERENT CLUSTERS)

$$s(P_M) = \min_{j=1\ldots M} s(C_j) = \min_{j,k,\ell \mid O_k \in C_j, O_\ell \notin C_j} d_{k\ell}$$

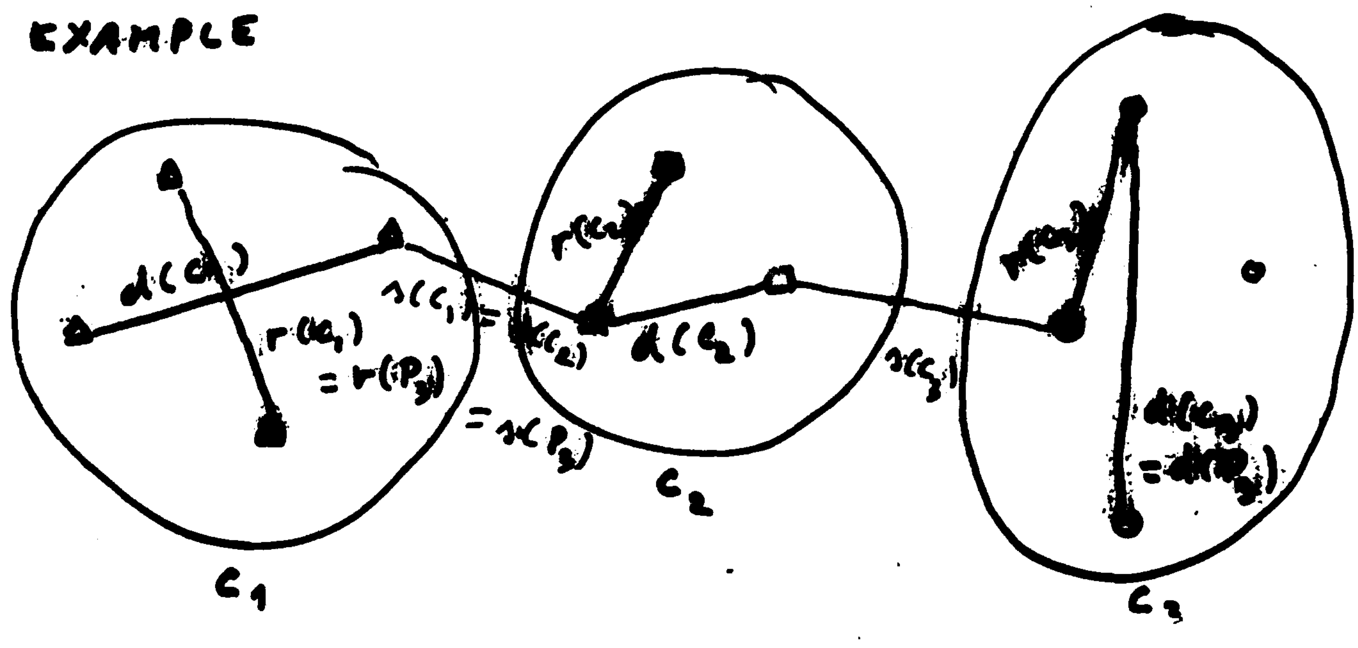
DIAMETER OF A PARTITION : MAXIMUM DIAMETER OF THIS PARTITION'S CLUSTERS (OR MAXIMUM DISSIMILARITY BETWEEN ENTITIES IN THE SAME CLUSTER

$$d(P_M) = \max_{j=1\ldots M} d(C_j) = \max_{j,k,\ell \mid O_k, O_\ell \in C_j} d_{k\ell}$$

RADIUS OF A PARTITION : MAXIMUM RADIUS OF THIS PARTITION'S CLUSTERS

$$r(P_M) = \max_{j=1,2,\ldots M} r(C_j) = \max_{j} \min_{k \mid O_k \in C_j} \max_{\ell \mid O_\ell \in C_j} d_{k\ell}$$

EXAMPLE



PROBLEM 1 FIND A MAXIMUM SPLIT PARTITION OF O INTO M CLUSTERS (SOLVED BEFORE POSED!

THEOREM 1 (P. ROSENSTIEHL, 67, REFORMULATED BY M. DELATTRE AND P.H. 1980).

ASSOCIATE A COMPLETE GRAPH $G=(V,E)$ WITH O ($v_j \leftrightarrow o_j$ AND WEIGHT ITS EDGES BY THE d_{kl}. THEN THE SPLIT VALUES FOR ALL CLUSTERS AND (HENCE) PARTITIONS OF O ARE EQUAL TO THE WEIGHTS OF THE EDGES OF A MINIMUM SPANNING TREE OF G.

COROLLARY 1 (M. DELATTRE, P.H. 1980) THE SINGLE LINKAGE ALGORITHM PROVIDES ONLY MAXIMUM SPLIT PARTITIONS.

SINGLE LINKAGE ALGORITHM, FIRST VERSION (JOHNSON 67, ZUBRISKI et al 58)

INITIALISATION $(O(N))$

$P_N = \{C_1, C_2, \ldots C_N\}$ WITH $C_j = \{o_j\}$ $j = 1 \ldots N$;

$M \leftarrow N-1$

CURRENT STEP $(O(N^2))$

WHILE $M > 0$ DO

FIND CLUSTERS C_j, C_k WITH ENDPOINTS OF SMALLEST DISSIMILARITY EDGE JOINING DIFFERENT CLUSTERS (≡ MINIMUM SPLIT)

$P_M \leftarrow [P_M \setminus \{C_j, C_k\}] \cup C_{2N-M}$ WITH $C_{2N-M} = C_j \cup C_k$

$M \leftarrow M-1$

EXAMPLE

$O = \{o_1, o_2, \ldots o_5\}$

$$D = \begin{vmatrix} - & 8 & 9 & 5 & 2 \\ 8 & - & 1 & 10 & 3 \\ 9 & 1 & - & 7 & 4 \\ 5 & 10 & 7 & - & 6 \\ 2 & 3 & 4 & 6 & - \end{vmatrix}$$

NOTE : THRESHOLD-TYPE CRITERIA ARE INVARIANT FOR A MONOTONE TRANSFORMATION OF THE DISSIMILARITIES

$P_5 = \{\{o_1\},\{o_2\},\{o_3\},\{o_4\},\{o_5\}\}$

$s(P_5) = 1 = d_{23}$

$P_4 = \{\{o_1\},\{o_2,o_3\},\{o_4\}\{o_5\}\}$

$s(P_4) = 2 = d_{15}$

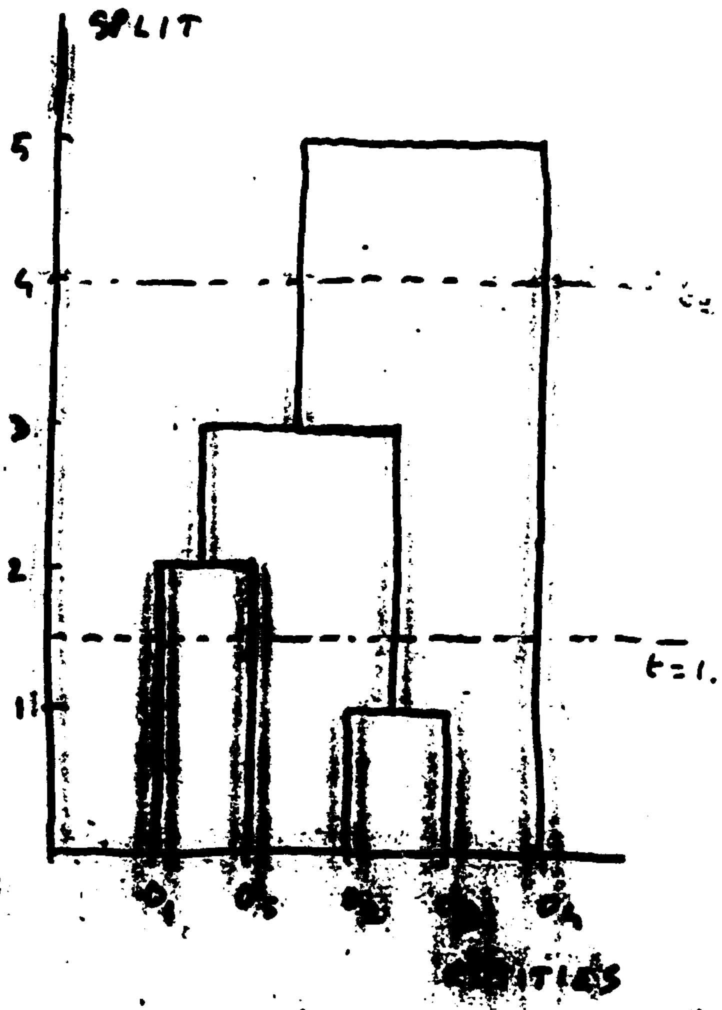
$P_3 = \{\{o_1,o_5\}\{o_2,o_3\}\{o_4\}\}$

$s(P_3) = 3 = d_{25}$

$P_2 = \{\{o_1,o_2,o_3,o_5\}\{o_4\}\}$

$s(P_2) = 5 = d_{14}$

$P_1 = \{\{o_1,o_2,o_3,o_4,o_5\}\}$



RESULTS PRESENTED ON A DENDROGRAM:

HORIZONTAL LINES : PARTITIONS

≡ CONNECTED COMPONENTS OF THRESHOLD GRAPH $G_t = (V, E_t)$ WITH $E_t = \{\{v_k, v_l\} \in E \mid d_{kl} \le t\}$

SINGLE LINKAGE ALGORITHM, SECOND VERSION
(GOWER, ROSS 1969)

INITIALISATION

COMPLEXITY:

FIND MST OF G — $O(N^2)$ WITH PRIM'S ALG

RANK EDGES OF MST BY INCREASING VALUES

$P_N = \{c_1, c_2, \dots c_N\}$ WITH $c_j = \{o_j\}$ $j = 1, \dots N$.

$n \leftarrow N-1$

CURRENT STEP

WHILE EDGES OF MST REMAIN DO

FIND CLUSTERS c_j, c_k CONTAINING ENDPOINTS OF TOP EDGE

$P_n \leftarrow [P_n \setminus (c_j, c_k)] \cup c_{2N-n}$ WITH $c_{2N-n} = c_j \cup c_k$

$n \leftarrow n-1$

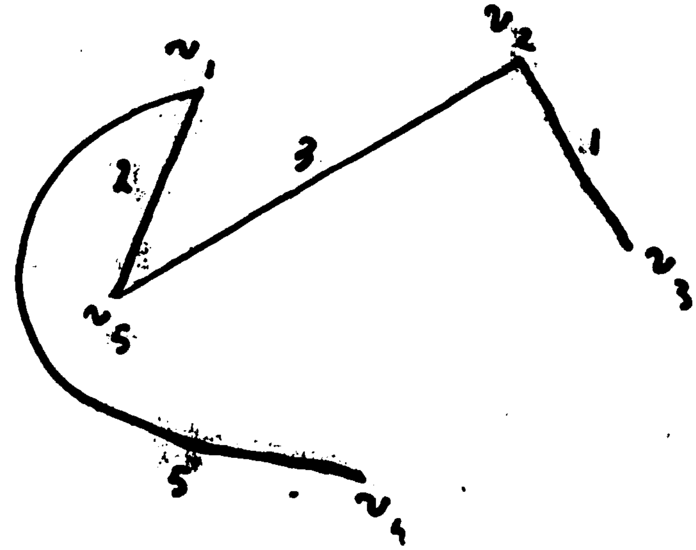
REMOVE EDGE — $O(N)$

ENDWHILE.

$\Rightarrow O(N^2)$ IN ALL.

BEST POSSIBLE ALGORITHM.

MST:



PROBLEM 2 FIND A MINIMUM DIAMETER PARTITION OF O INTO M CLUSTERS

THEOREM 2 (BRUCKER, 78, P.H., M DELATTRE, 78)
MINIMUM DIAMETER PARTITIONING IS NP-HARD FOR $M \geq 3$.

PROOF: REDUCTION TO 3-COLORING.

ALGORITHM BASED ON GRAPH COLORING:
RAO 71 (FOR $M=2$), CHRISTOFIDES 75 (PRINCIPLE)
P.H., M. DELATTRE 78.

INITIALISATION

RANK EDGES BY NON-INCREASING d_{kl}

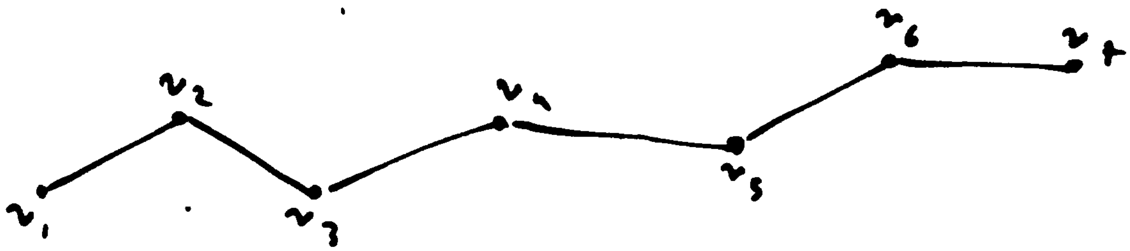
CONSIDER THRESHOLD GRAPHS G_t WITH $t = d_{kl}$ IN ORDER OF RANKING (FIRST G_t HAS NO EDGES, ALL VERTICES OF SAME COLOR)

CURRENT STEP

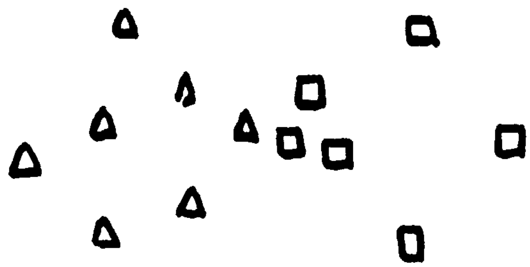
COLOR G_t; IF $\leq M$ COLORS NEEDED CONTINUE;
LAST GRAPH COLORABLE IN M COLORS HAS MINIMUM DIAMETER.

4. BICRITERION CLUSTER ANALYSIS (P.H., M.D., M.M
IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE

- CHAINING EFFECT : WHEN USING S.L.A
VERY DISSIMILAR ENTITIES AT THE ENDS OF CHAIN OF PAIRWISE SIMILAR ENTITIES IN SAME CLUSTER



DISSECTION EFFECT : WHEN USING C.L.A.
- VERY SIMILAR ENTITIES IN DIFFERENT CLUSTERS



COMPROMISE AVOIDING BOTH DEFECTS : FIND EFFICIENT SOLUTIONS P_M FOR SPLIT AND DIAMETER :

- $\nexists P'_M \mid s(P'_M) > s(P_M)$ AND $d(P'_M) \leq d(P_M)$
OR $s(P'_M) \geq s(P_M)$ AND $d(P'_M) < d(P_M)$

PRINCIPLE OF ALGORITHM: FIND MST T OF G; IMPOSE MINIMUM VALUE FOR SPLIT BY FUSIONING VERTICES AT EXTREMITIES OF SMALLEST EDGE OF T, SECOND SMALLEST ...

IN PRACTICE: WORKS FOR 70-100 ENTITIES MORE IF THE IS SOME STRUCTURE.

ANALYSIS OF RESULTS THROUGH THE DIAMETER-SPLIT MAP (OR QUALITIES MAP

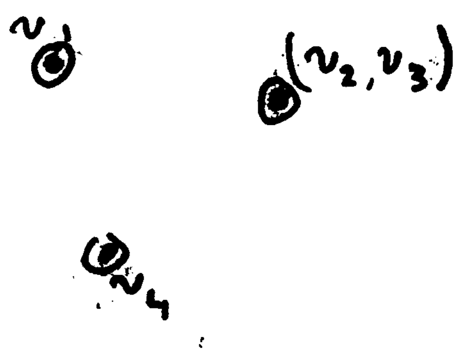
INDICATIONS OF STRUCTURE:

- SINGLE EFFICIENT PARTITION FOR SOME M
- STRONG DECREASE IN DIAMETER FOR SMALL DECREASE IN SPLIT,
- STRONG DECREASE IN DIAMETER FOR UNIT INCREASE IN M.

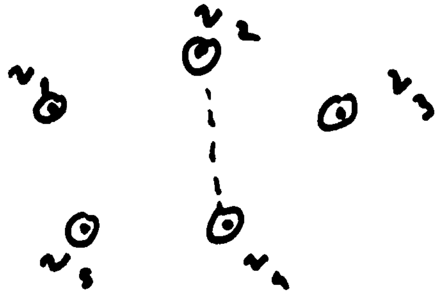
EX

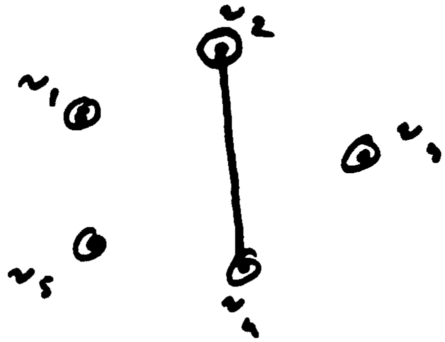
$$D = \begin{vmatrix} 0 & 2 & 5 & 6 \\ 2 & 0 & 1 & 3 \\ 5 & 1 & 0 & 4 \\ 6 & 3 & 4 & 0 \end{vmatrix}$$

G^1

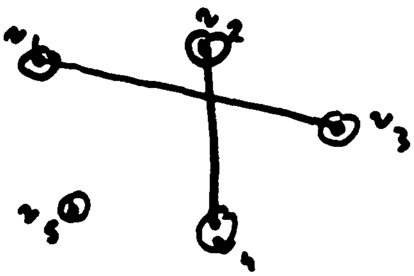


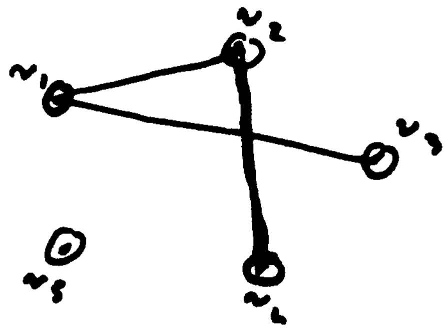
EXAMPLE.

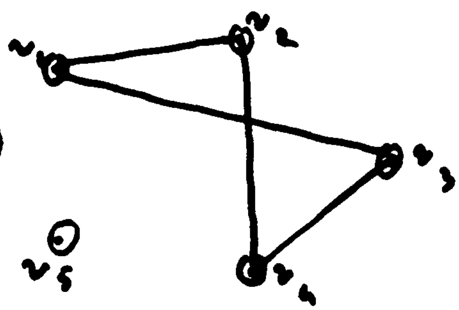


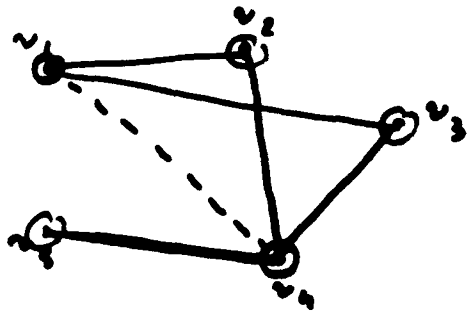


$d(P_1) = d_{24} = 10$









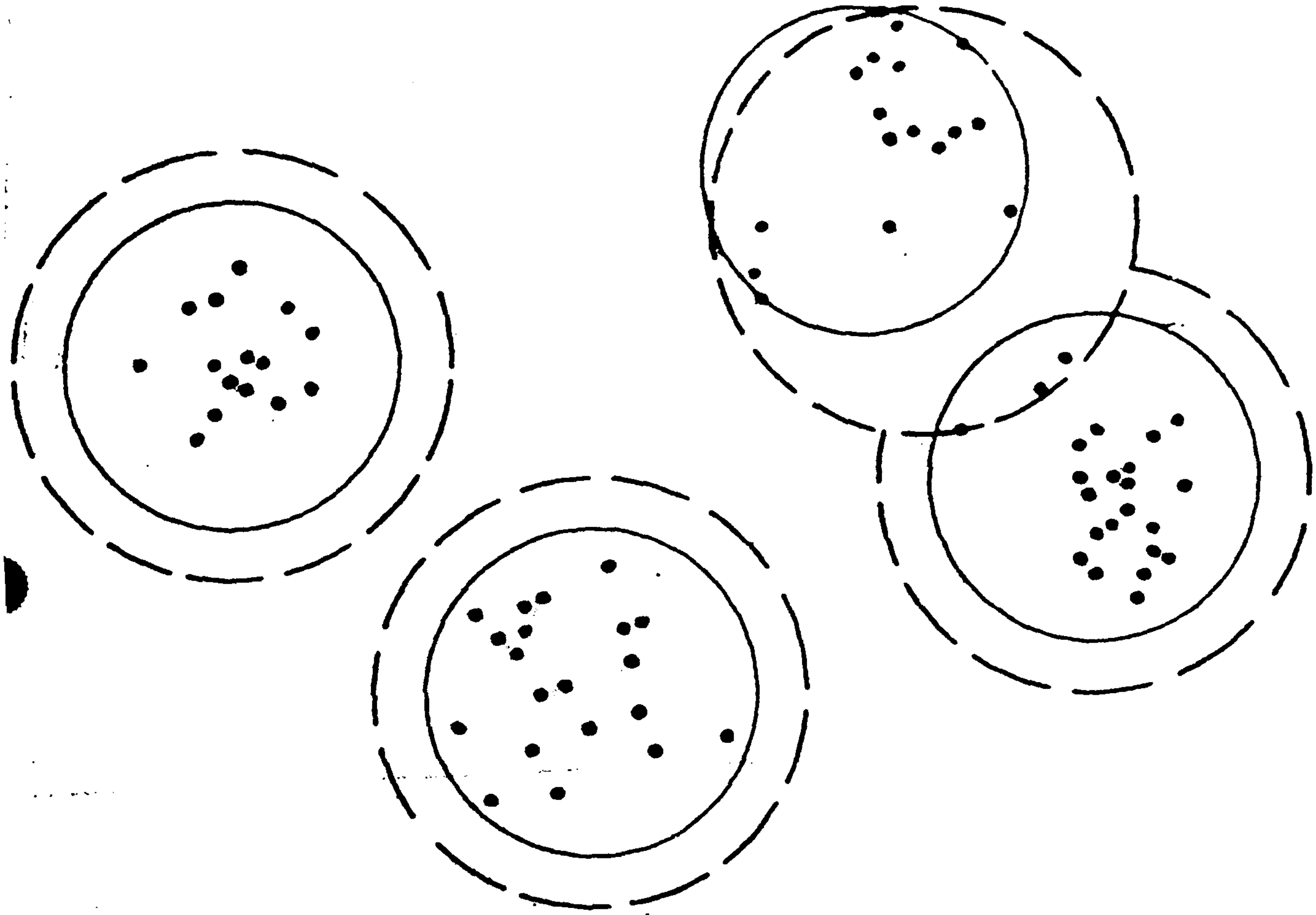
$d(P_2) = 5$.

$P_2 = \{\{o_1, o_4\}, \{o_2, o_3, o_5\}\}$

IN PRACTICE: WITH STRUCTURE, LARGE INSTANCES (N = 600, 15 COLORS)

WITHOUT STRUCTURE $N \geq 75-100$

C. Ruspini's Data: Partitions Into Four Groups.



—— : CLUSTERGRAPH 1 (A.H.S)

— — : HCLUSTER (G.C.)

PROBLEM 3 FIND A MINIMUM RADIUS PARTITION OF O INTO M CLUSTERS

- LITTLE STUDIED IN CLUSTERING (KRIVANEK SHOWS IT IS NP-HARD FOR $M \geq 3$), KNOWN IN LOCATION THEORY AS THE P-CENTER PROBLEM (MINIEKA 70, KARIV-HAKIMI 79, DREZNER 84 ...)
  PRINCIPLE OF ALGORITHM: PRUZAN-KRARUP, 82

INITIALISATION

RANK EDGES BY NON-INCREASING d_{kl}.

DICHOTOMOUS SEARCH FOR OPTIMAL VALUE;

TENTATIVE VALUE t

CURRENT STEP

- SOLVE COVERING PROBLEM

$$\min z = \sum_{k=1}^{N} x_k$$

$$\text{s.t.} \quad \sum_{k=1}^{N} a_{ik} x_k \geq 1 \qquad i = 1, 2, \dots N$$

$$x_k \in \{0, 1\} \qquad k = 1, 2, \dots N$$

WHERE $a_{ik} = \begin{cases} 1 & \text{IF } d_{ik} \leq t \\ 0 & \text{OTHERWISE} \end{cases}$

IF $z^* > M$ INCREASE t; OTHERWISE DECREASE t

STOP WHEN WHOLE RANGE COVERED.

- $\Rightarrow O(\log N)$ COVERING PROBLEMS (WHICH ARE POLYNOMIAL FOR FIXED M (DREZNER 84))

Figure 3 : Diameter split-map - Uniform data

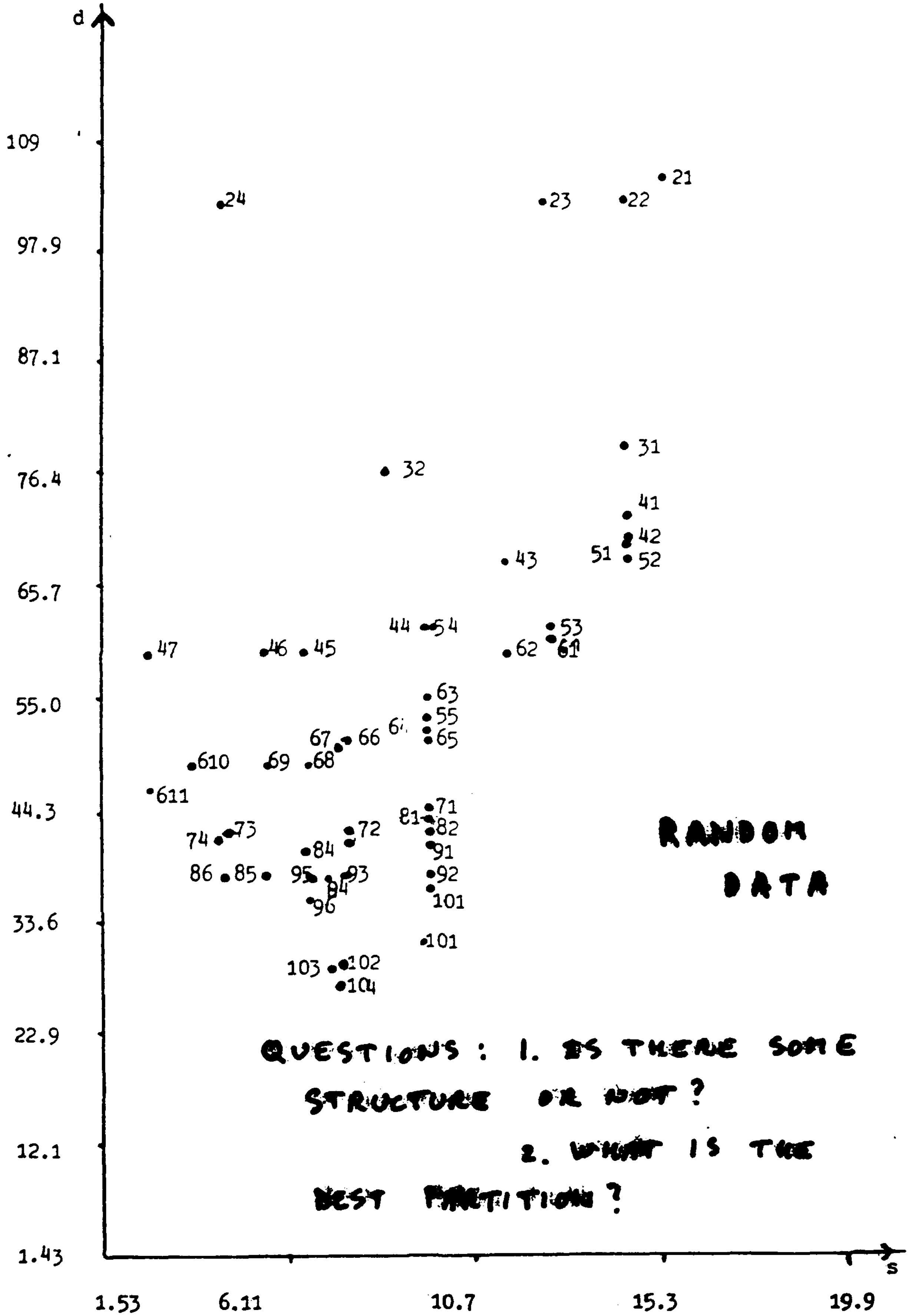
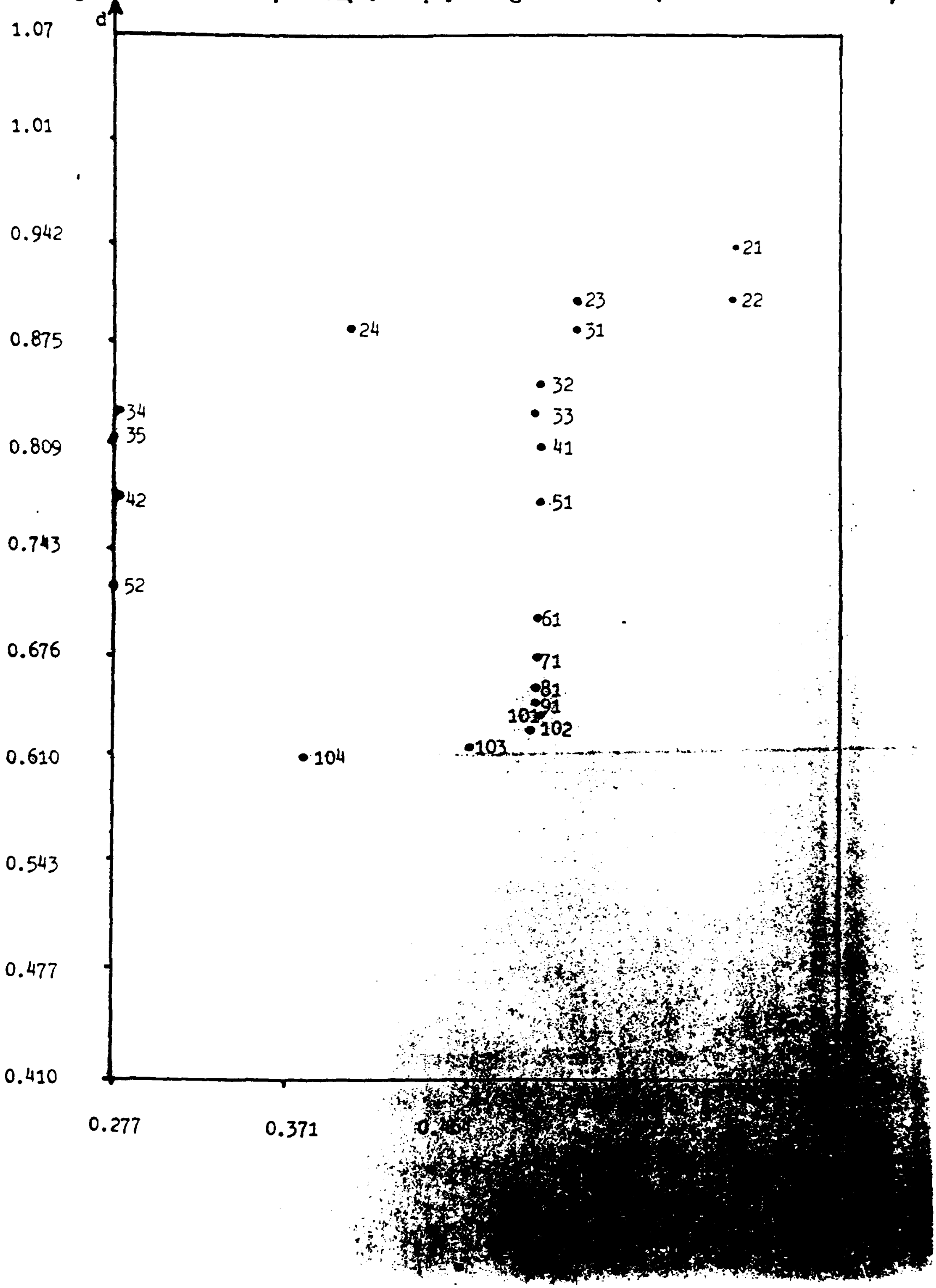


Table 2. 24 psychological tests (HARMAN, 1964)

Test	B-coeff. partition	61 bicrit. partition	Factor associated
1. Visual Perception	1	1	
2. Cubes	1	1	Spatial
3. Paper Form Board	1	1	Relations
4. Flags	1	2	
5. General Information	2	3	
6. Paragraph Comprehension	2	3	
7. Sentence Completion	2	3	Verbal
8. Word Classification	2	3	
9. Word Meaning	2	3	
10. Addition	3	4	
11. Code	3	4	Perceptual
12. Counting Dots	3	4	Speed
13. Straight-Curved Capitals	3	4	
14. Word Recognition	4	5	
15. Number Recognition	4	5	
16. Figure Recognition	4	5	Recognition
17. Object-Number	4	5	Associative
18. Number-Figure	4	6	Memory
19. Figure-Word	4	6	
20. Deduction	5	3	
21. Numerical Puzzles	5	[illegible]	
22. Problem Reasoning	5	5	
23. Series Completion	5	[illegible]	
24. Arithmetic Problems	5	[illegible]	

Figure 1. Diameter-split map : 24 psychological tests (HARMAN'S DATA)



SOME CONSTRAINED CLUSTERING PROBLEMS WITH THRESHOLD TYPE CRITERIA

PROBLEM 5 FIND A WEIGHT CONSTRAINED MAXIMUM SPLIT PARTITION OF O INTO M CLUSTERS (ENTITY O_i HAS WEIGHT w_i; MAXIMUM WEIGHT FOR A CLUSTER = $\bar{w}$)

NP-HARD PROBLEM (REDUCTION TO 3-PARTITION)

PRINCIPLE OF ALGORITHM: DICHOTOMOUS SEARCH ON VALUE OF SPLIT; FOR A GIVEN VALUE t THE EDGES OF THE MST WITH $d_{kl} < t$ INDUCE A SPANNING FOREST OF G; EXISTENCE OF A SOLUTION ≡ BIN-PACKING PROBLEM. EFFICIENT ALGORITHM FOR THE LATTER (MARTELLO, TOTH 89, 90) ALLOWS TO SOLVE PROBLEMS WITH N = 500 (P.H, A.JAUMARD, K MUSITU 1990)

PROBLEM 6 FIND A CONNECTIVITY CONSTRAINED MAXIMUM SPLIT PARTITION OF O INTO M CLUSTERS

NP-HARD (REDUCTION TO STEINER'S PROBLEM ON GRAPHS)

ONGOING WORK WITH B. SIMEONE.

3. THRESHOLD - SUM TYPE CRITERIA

SUM-OF-SPLITS OF A PARTITION

$$ss(P_M) = \sum_{j=1,\dots,M} s(C_j) = \sum_{j=1,\dots,M} \min_{k,l \mid O_k \in C_j, O_l \notin C_j} d_{kl}$$

SUM-OF-DIAMETERS OF A PARTITION

$$sd(P_M) = \sum_{j=1..M} d(C_j) = \sum_{j=1..M} \max_{k,l \mid O_k, O_l \in C_j} d_{kl}$$

SUM-OF-RADII OF A PARTITION

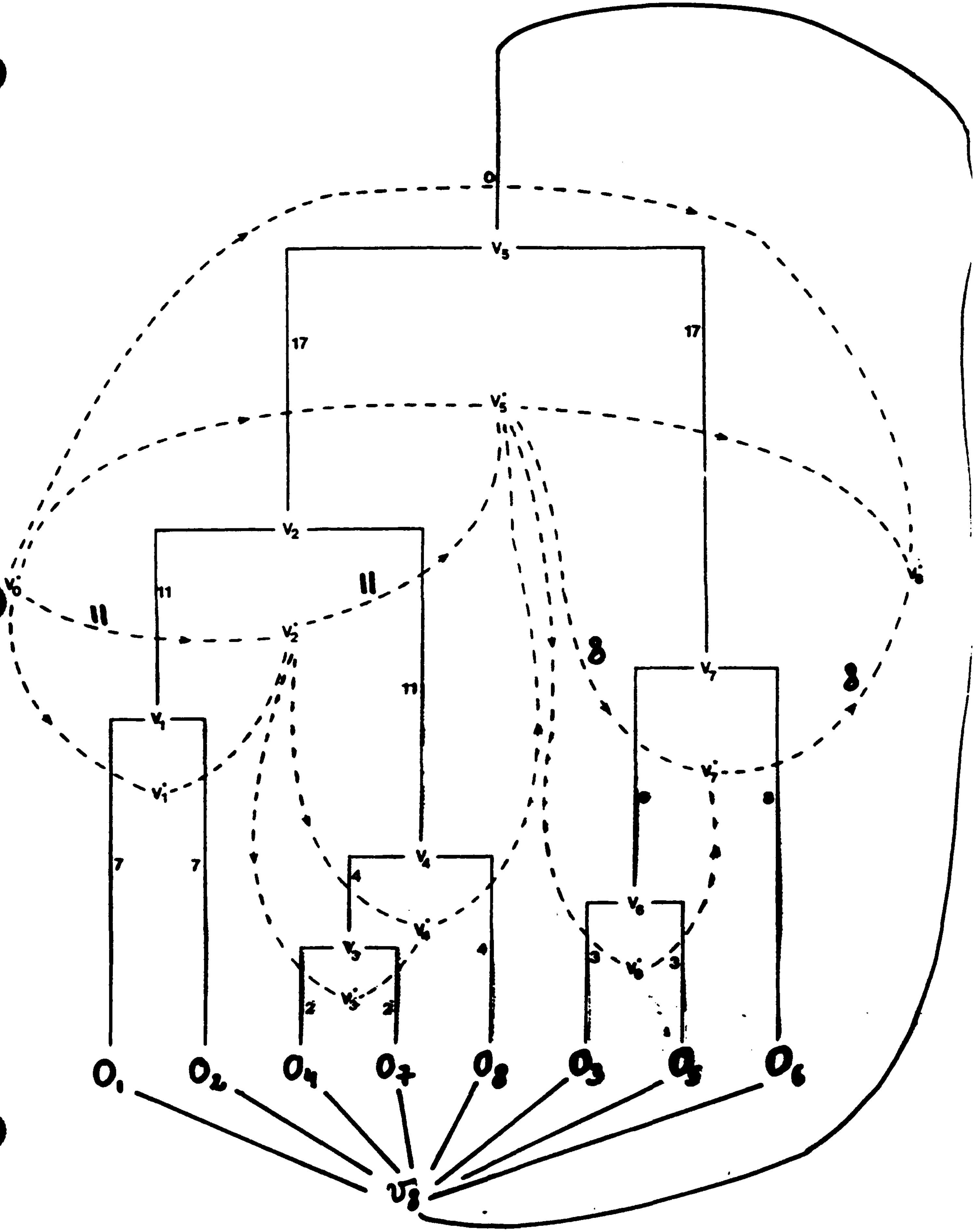
$$sr(P_M) = \sum_{j=1..M} r(C_j) = \sum_{j=1..M} \min_{k \mid O_k \in C_j} \max_{l \mid O_l \in C_j} d_{kl}$$

PROBLEM ? FIND A MAXIMUM SUM-OF-SPLITS PARTITION OF O INTO M CLUSTERS

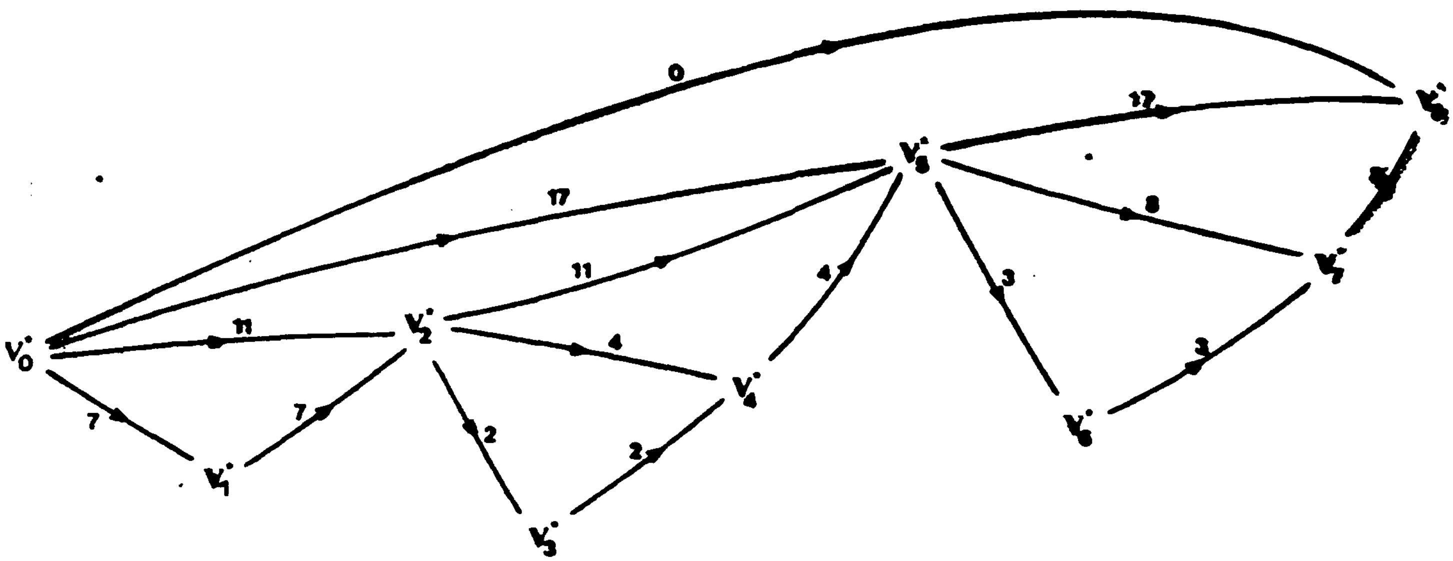
THEOREM (P.H., B. JAUMARD, O. FRANK, 89) THERE IS AN OPTIMAL PARTITION FOR THE MAXIMUM SUM-OF-SPLITS CRITERION WHICH CONTAINS ONLY CLUSTERS OBTAINED FOR THE MAXIMUM SPLIT CRITERION i.e. BY THE SINGLE LINKAGE ALGORITHM

THE PROBLEM CAN BE SOLVED IN $O(N^2)$ TIME USING THE DUAL GRAPH OF THE SINGLE-LINKAGE DENDROGRAM

DUAL GRAPH OF A DENDROGRAM



G_D^* is acyclic, N+1 vertices, 2N-1 arcs



The maximum sum-of-splits partitions P_M for $M = 2, 3, \ldots, N-1$, correspond to the longest paths between the first and last vertices of the dual graph G_D^* containing $2, 3, \ldots, N-1$ arcs respectively.

PROBLEM 8 FIND A MINIMUM SUM OF DIAMETER PARTITION OF O INTO M CLUSTERS

THEOREM (BRUCKER, 78) NP-HARD FOR $M \geq 3$. (CASE $M=2$ OPEN)

ALGORITHM USING BOOLEAN ALGEBRA FOR $M=2$ (P.H, B. JAUMARD, 87)

SUBPROBLEM: GIVEN DIAMETERS $d_1 (\geq) d_2$ IS THERE A SOLUTION?

$$x_k = \begin{cases} 1 & O_k \text{ IN } C_1 \\ 0 & O_k \text{ IN } C_2 \end{cases}$$

$$\Rightarrow (1-x_k)(1-x_\ell)=0 \quad \forall\, d_{k\ell} > d_2$$

$$x_k x_\ell + (1-x_k)(1-x_\ell) = 0 \quad \forall\, d_{k\ell} > d_1$$

QUADRATIC LOGICAL EQUATION, SOLVABLE IN $O(N^2)$ TIME BY ASPVALL, PLASS, TARJAN ALGO

AS N^2 POSSIBLE VALUES FOR $d_1, d_2 \Rightarrow O(N^6)$

THEOREM (P.H., B. JAUMARD, 87) THE DIAMETER OF ANY BIPARTITION OF O IS EQUAL TO THE WEIGHT OF THE LARGEST EDGE CLOSING AN ODD CYCLE WITH THE MAXIMUM SPANNING TREE T OF G, OR TO THE WEIGHT OF AN EDGE OF T LARGER THAN IT $\Rightarrow O(N)$ CANDIDATE VALUES FOR d_1

WITH DICHOTOMOUS SEARCH $\Rightarrow O(N^3 \log N)$ ALGORITHM.

PROBLEM 9 FIND A MINIMUM SUM-OF-RADII PARTITION OF O INTO n CLUSTERS (P-CENT SUM PROBLEM)

NP-HARD (P.H., M. LABBE, M. MINOUX)

REDUCTION TO DOMINATING SET

SOLUTION BY MIXED-INTEGER PROGRAMMING (EXTENSION OF COLUMN GENERATION TECHNIQUE TO THE MIXED-INTEGER CASE)

ALSO OF INTEREST: CONTINUOUS CASE VERSION: CENTERS NOT AMONG AN A PRIORI SET BUT IN THE PLANE; DISSIMILARITIES EQUAL TO EUCLIDEAN DISTANCES

M-CENTER SUM

$$\text{MIN} \quad \sum_{i=1}^{m} z_i$$

$$\text{s.t.} \quad z_i \geq d_{ij}\, x_{ij} \qquad i = 1,2,\ldots,m \quad j = 1,2,\ldots,n,\ j \neq i$$

$$\sum_{j=1}^{n} x_{ij} = 1 \qquad i = 1,2,\ldots,m$$

$$-x_{ij} + y_i \geq 0 \qquad i = 1,2,\ldots,m \quad j = 1,2,\ldots,n$$

$$\sum_{i=1}^{m} y_i = M$$

$$x_{ij},\, y_i \in \{0,1\} \qquad i = 1,2,\ldots m \quad j = 1,2,\ldots n$$

LINEAR RELAXATION [illegible] YIELD VERY [illegible] BOUND?

ALTERNATE FORMULATION :

$$\text{MIN} \quad \sum_{k} c_k x_k$$

$$\text{s.t.} \quad \sum_{k} a_{ik} x_k \geq 1$$

$$\sum_{k} x_k = M$$

$$x_k \in \{0,1\}$$

COVERING PROBLEM WITH EXTRA CONSTRAINT

A^k : $a_{ik} = 1$ VERTEX i IN SET k.
$a_{ik} = 0$ OTHERWISE

4. SUM TYPE CRITERIA

STAR OF A CLUSTER C_j : MINIMUM OVER ALL ENTITIES OF THAT CLUSTER OF THE SUM OF ALL DISSIMILARITIES BETWEEN THAT ENTITY (THE MEDIAN) AND ALL OTHERS IN THE CLUSTER.

$$st(C_j) = \min_{k \mid O_k \in C_j} \sum_{l \mid O_l \in C_j} d_{kl}$$

CLIQUE OF A CLUSTER C_j : SUM OF ALL DISSIMILARITIES BETWEEN PAIRS OF ENTITIES IN THAT CLUSTER

$$cl(C_j) = \sum_{k,l \mid O_k, O_l \in C_j} d_{kl}$$

VARIANCE ....

TYPICAL PROBLEM : FIND PARTITION OF O INTO M CLUSTERS SUCH THAT THE LARGEST $\left\{ \begin{array}{l} \text{STAR} \\ \text{CLIQUE} \end{array} \right\}$ IS MINIMUM

OPEN PROBLEMS UP TO NOW

PROBLEM 10 : FIND A PARTITION OF O INT[O]
m CLUSTERS WITH MINIMUM SUM-OF-STARS
≡ THE P-MEDIAN PROBLEM
ERLENKOTTER 78 , KORKEL 89
IN CLUSTERING : MULVEY 82
ALSO VERSION WITH CAPACITY (SIZE) CONSTRAI[NTS]
· MULVEY , BECK

PROBLEM 11 : FIND A PARTITION OF O
(INTO m CLUSTERS) WITH MINIMUM
SUM-OF-CLIQUES (OR WITH MINIMUM
WITHIN CLUSTERS SUM OF DISSIMILARITIES)
CLOSELY RELATED PROBLEM WITH d_{kl} ∈ [R]
(SOME NEGATIVE VALUES) m UNFIXED : MEDIAN RELATION
MICHAUD , MARCOTORCHINO (1981) TRIANGL[E]
INEQUALITIES
GROTSCHEL, WAKABAYASHI (1989, 1990) :
CUTTING PLANE ALGORITHM .
FOR FIXED m : INTEGER PROGRAMMING
KLEIN , ARONSON (1991)

CONCLUSIONS

TRADITIONAL METHODS OF CLUSTER ANALYSIS:

MATHEMATICALLY VAGUE: NO EXPLICIT CRITERION

FEW PROPERTIES

RECENT MATHEMATICAL ADVANCES

- MATHEMATICAL PROGRAMMING

  (VINOD 69, RAO, 71, ...)
  (GROTSCHEL, WAKABAYASHI 89, 90, CONFORTI, RAO

- GRAPH THEORY

  (MATULA, P.H. ...)

- ORDER THEORY

  (JANOWITZ, BARTHELEMY, LECLERC, MONJARDET..

MATH. PROG. + GRAPH THEORY HAS ALLOWED TO

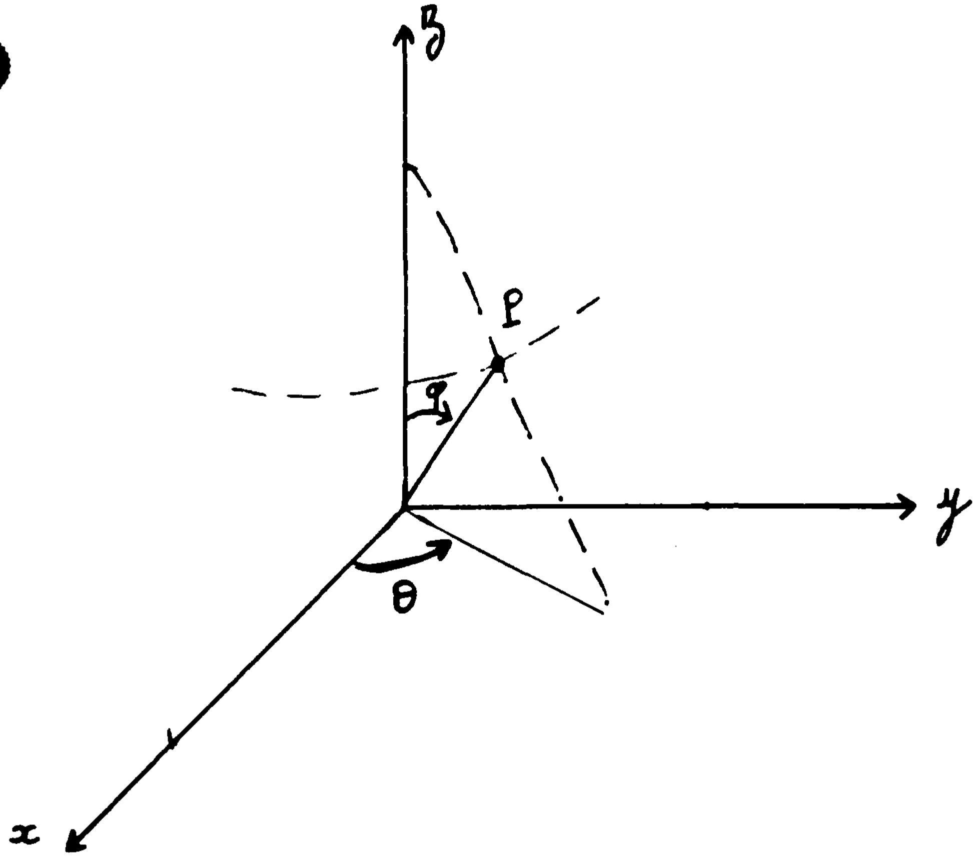
- FORMULATE PRECISE, SOMETIMES NEW PROBLEMS
- STUDY THEIR COMPLEXITY
- OBTAIN EFFICIENT NEW ALGORITHMS, SOMETIMES BEST POSSIBLE
- MAKE CONNECTIONS WITH OTHER CHAPTERS OF MATH. PROG., OPERATIONS RESEARCH AND GRAPH THEORY.

FACILITY LOCATION PROBLEMS ON A SPHERE

Pierre Hansen, HEC-GERAD, Montréal

Brigitte Jaumard, GERAD & Ecole Polytechnique de Montréal

Stéphane Krau, Ecole Polytechnique de Montréal
Ph.D. Student



φ = latitude

θ = longitude

$-\frac{\pi}{2} \leq \varphi \leq \frac{\pi}{2}$

$-\pi \leq \theta \leq \pi$

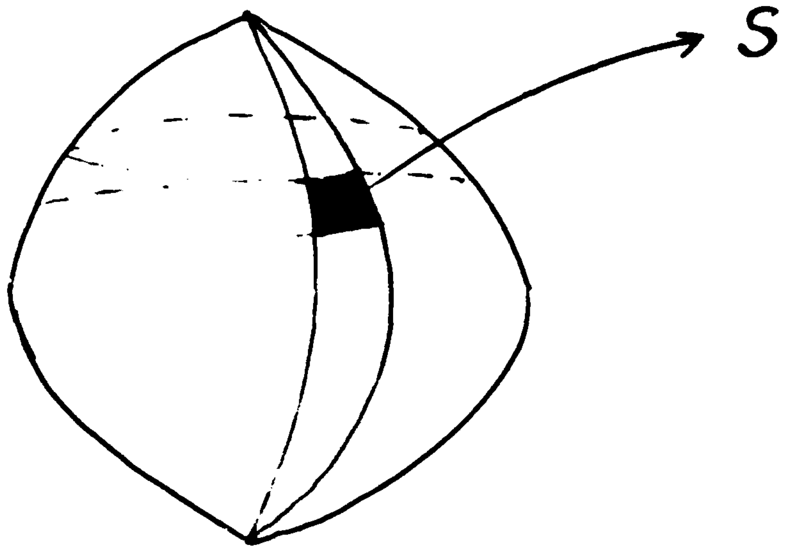
n given points $\quad M_i\ (\varphi_i, \theta_i) \quad i = 1, 2, \ldots, n$

Find $M(\varphi, \theta)$ such that :

$$\min_{\varphi, \theta} f(\varphi, \theta) = \sum_{i=1}^{n} w_i\, d(M, M_i)$$

$$\cos\left(d(M, M_i)\right) = \cos\varphi \cos\varphi_i \cos(\theta - \theta_i) + \sin\varphi \sin\varphi_i$$

Big Square Small Square Method (Hansen et al., 8.



"Square": Intersection area of 2 longitudes and 2 latitudes

Lower Bound: $g(S) = \sum_{i=1}^{n} \mathrm{dist}(M_i, S)$

$$\min_{S \in \mathcal{S}} g(S) \leq \min_{\varphi,\theta} f(\varphi,\theta)$$

Upper Bound: M_S midpoint of S : $M_S(\varphi^S, \theta^S)$

$$\min_{\varphi,\theta} f(\varphi,\theta) = \min_{S \in \mathcal{S}} f(\varphi^S, \theta^S)$$

(Dhar et Rao, 82)

A ship has to be located in a specific region in Philippine Basin within the quadrilateral formed by

$$137.00^\circ \leq \theta < 150.00^\circ$$

$$12.50^\circ \leq \varphi < 18.50^\circ$$

and must be ready to intervene in case of need, at any one of 15 given ports in the eastern hemisphere. Needs may occur according to available forecasts which are used as weights. Assuming the cost as a linear function of geodetic distance, the objective is to minimize weighted average of cruise distances to the various ports.

cpu 4.52 sec. on a SPARC 2.

objective = 612.2417

$\varphi^* = 21.621$ $\theta^* = 120.059$

Ports and their locations

Ports	(φ, θ) in degrees	weights
Mapotu	$(-26.00, 32.41)$	1
Aden	$(12.75, 45.20)$	1
Mogadishu	$(2.02, 45.33)$	1.5
Abadan	$(30.33, 48.27)$	0.5
Muscat	$(23.62, 58.58)$	0.5
Christmas Is.	$(-10.50, 105.66)$	1
Saigon	$(10.77, 106.67)$	1
Perth	$(-31.93, 115.83)$	0.3
Manila	$(14.58, 120.98)$	1
Taipei	$(25.00, 121.50)$	1.5
Inchion	$(37.46, 126.63)$	2
Tokyo	$(35.70, 139.77)$	1
Guam	$(13.47, 144.78)$	2
Port Moresby	$(-9.5, 147.16)$	1
Wake Is.	$(19.32, 166.60)$	1

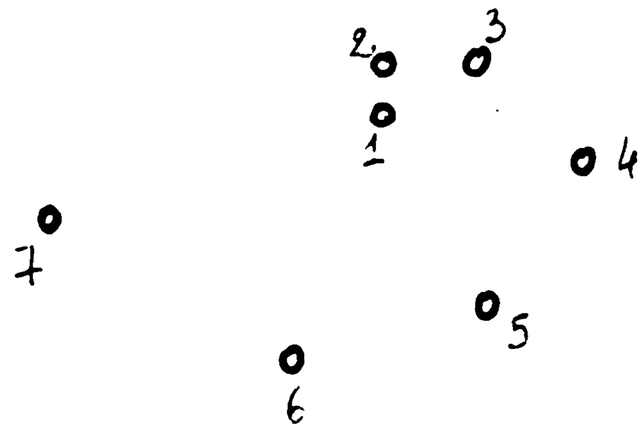
Pierre HANSEN, GERAD-HEC, Montréal

Brigitte JAUMARD, Ecole Polytechnique de Montréal

Bruno SIMEONE, University of Rome

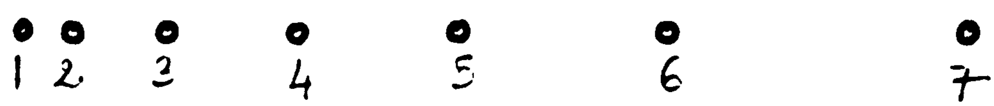
Martin KPOMASSE, MsC. Student Ecole Polytechnique de Montréal

dissimilarity = euclidean distance.

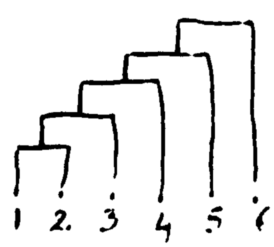


Sample 1

Sample 2



Same Dendrograms !



$O = \{O_1, O_2, O_3, O_4, O_5, O_6, O_7, O_8, O_9\}$

- Dissimilarity Matrix.

	1	2	3	4	5	6	7	8	9
1									
2	12.								
3	65	10							
4	86	80	93						
5	72	70	53	66					
6	79	73	56	45	54				
7	7	97	67	33	42	51			
8	49	32	70	15	5	56	66		
9	82	80	74	73	95	74	23	46	

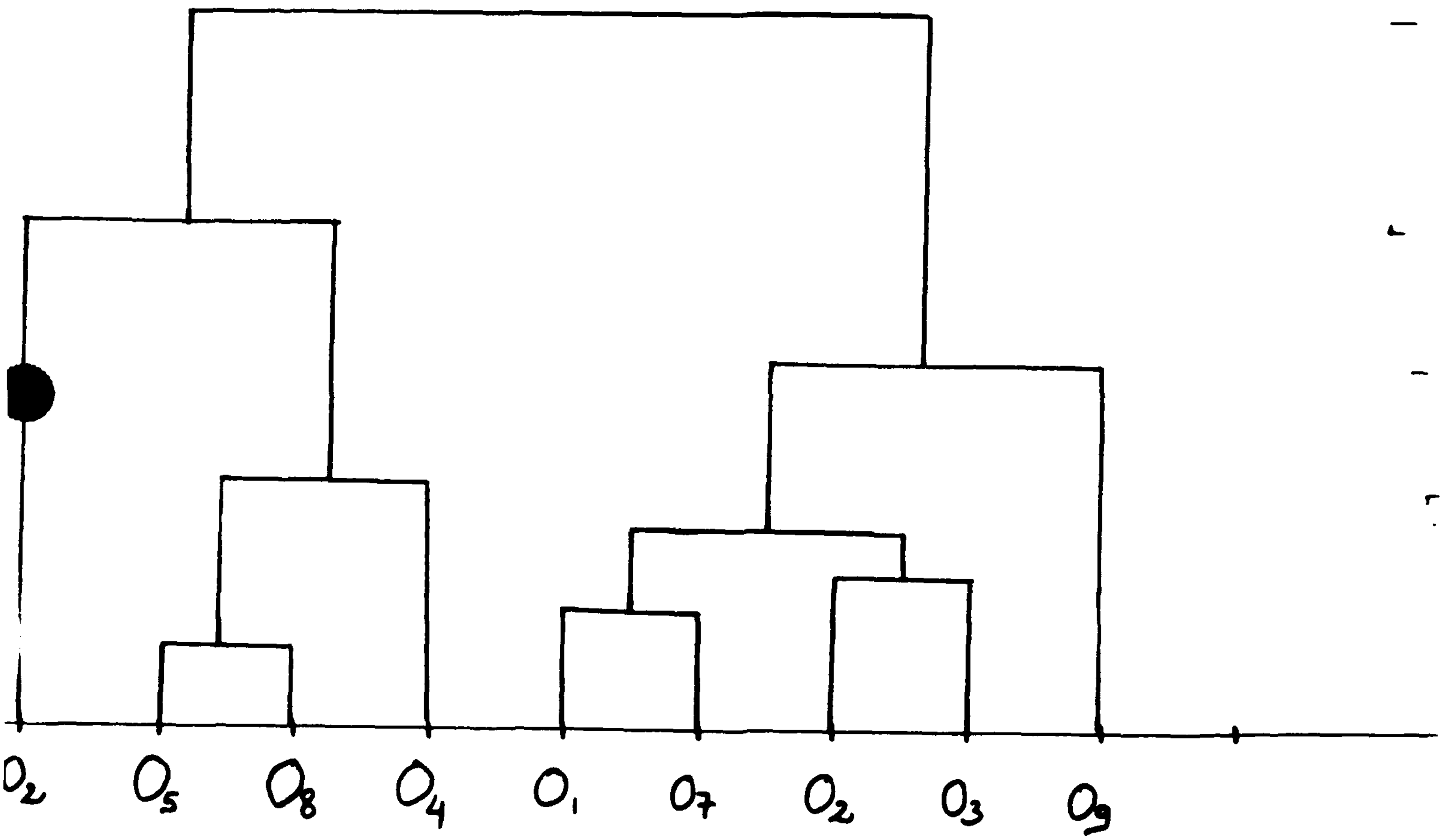
- $P_1 = \{\{O_1\}, \{O_2\}, \{O_3\}, \{O_4\}, \{O_5\}, \{O_6\}, \{O_7\}, \{O_8\}, \{O_9\}\}$

$P_2 = \{\{O_1\}; \{O_2\}; \{O_3\}; \{O_4\}; \{O_5, O_8\}; \{O_6\}; \{O_7\}; \{O_9\}\}$

$P_3 = \{\{O_1, O_7\}; \{O_2\}; \{O_3\}; \{O_4\}; \{O_5, O_8\}; \{O_6\}; \{O_9\}\}$

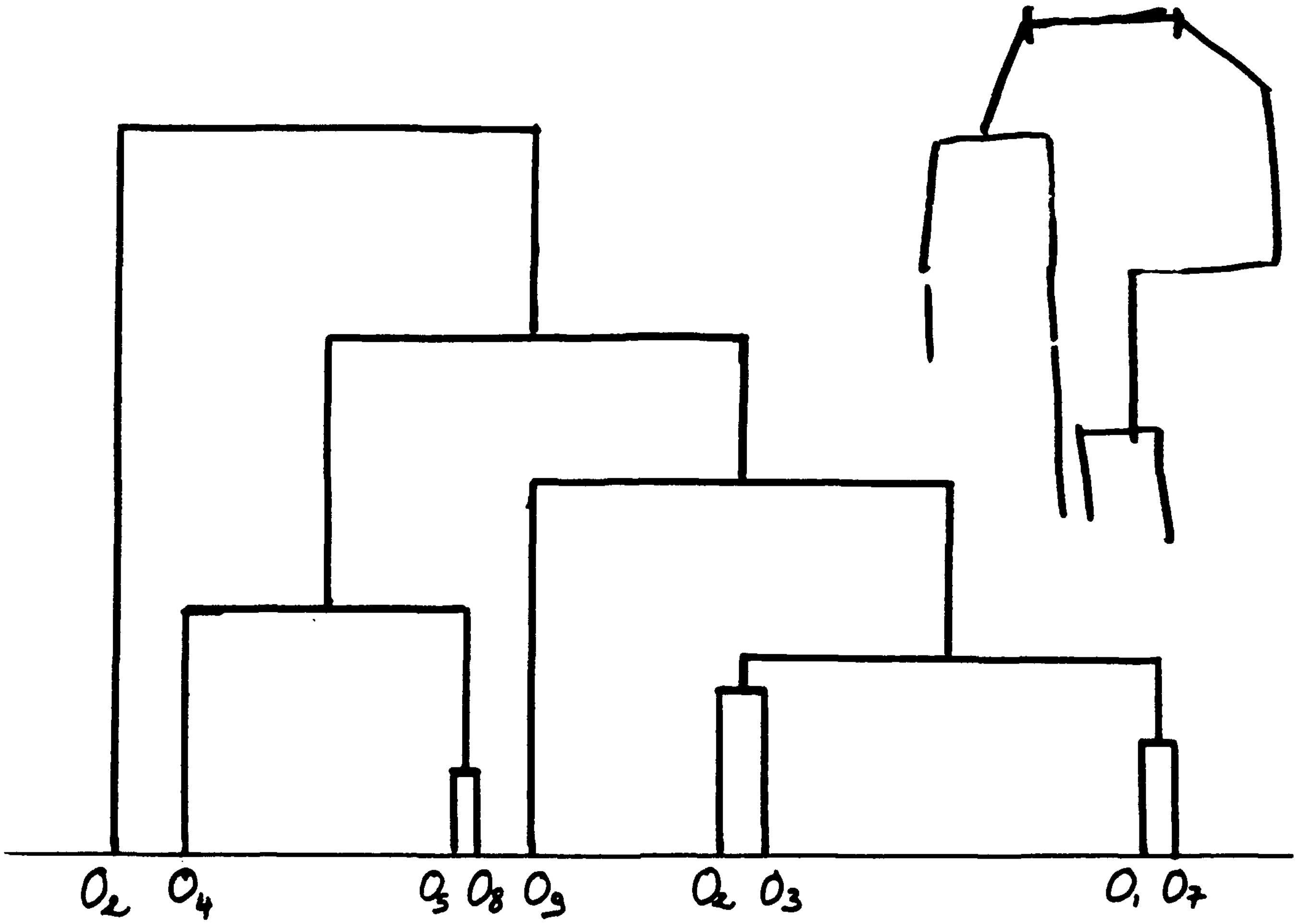
$P_4 = \{\{O_1, O_7\}; \{O_2, O_3\}; \{O_4\}; \{O_5, O_8\}; \{O_6\}; \{O_9\}\}$

$P_5 = \dots$



Good representation of the split

What about the diameter



Vertical lines : Split

Horizontal lines : Diameter

AVERAGE-LINKAGE DIVISIVE HIERARCHICAL CLUSTERING

PIERRE HANSEN

GERAD, ECOLE DES H.E.C., MONTREAL

BRIGITTE JAUMARD

GERAD, ECOLE POLYTECHNIQUE, MONTREAL

ELISABETE DA SILVA

ECOLE POLYTECHNIQUE, MONTREAL

$O = \{O_1, O_2, \dots, O_N\}$ set of N entities

$D = (d_{k\ell})$ matrix of dissimilarities :

$d_{k\ell} \geqslant 0 \qquad d_{k\ell} = d_{\ell k} \qquad d_{kk} = 0 \qquad \forall k, \ell$

$P_M = \{C_1, C_2, \dots, C_M\}$ partition of O into M clusters

$C_j \neq \emptyset \qquad C_i \cap C_j = \emptyset \qquad \bigcup_{j=1}^{M} C_j = O \qquad \forall i, j$

$\Pi_2(C_j)$ = set of bipartitions of C_j

H = hierarchy

$$C_i \in P_k \qquad C_j \in P_\ell \qquad k > \ell \Rightarrow \begin{cases} \text{either} & C_i \cap C_j = \emptyset \\ \text{or} & C_i \subset C_j \end{cases}$$

Divisive Hierarchical Clustering Algorithm

Input N, D ;

$P_1 = \{C_1\} = \{O\}$; $M = 1$;

<u>While</u> $M < N$ <u>do</u>

Select $C_j \in P_M$ according to Criterion 1 ;

Determine $\{C_{2M}, C_{2M+1}\} \in \Pi_2(C_j)$ according to Criterion 2 ;

$P_M \leftarrow (P_M \cup \{C_{2M}, C_{2M+1}\}) \setminus \{C_j\}$;

$M \leftarrow M + 1$

<u>End While</u>

Average - linkage :

Criterion 1 : maximization of the average dissimilarity of the best bipartition for all clusters:

Choose C_j such that :

$$\bar{d}^*(C_j) = \max_{i=1,2,\dots,M} \bar{d}^*(C_i)$$

Criterion 2: maximization of the average dissimilari $\bar{d}$ between pairs of entities, one in C_{2M}, another one in C_{2M+1}.

Introduction of binary variables x_r associated with each entity $O_r \in C_j$ such that

$$x_r = \begin{cases} 1 & \text{if } O_r \in C_{2M} \\ 0 & \text{otherwise} \end{cases}$$

Finding $\bar{d}^*(C_j)$ is then equivalent to:

$$\max_x \bar{d}(x) = \frac{\sum_{r \in R_j} \sum_{s \in R_j} d_{rs}\, x_r (1-x_s)}{\sum_{r \in R_j} x_r \sum_{r \in R_j} (1-x_r)}$$

$t_q \leftarrow t_q - 1$ FOR ALL q SUCH THAT $t_q > 0$

END REPEAT

END WHILE.

IN ORDER TO ACCELERATE ALGORITHM:

UPDATING OF CHANGES:

LET $\bar{x}_q = 1 - x_q$

$$\delta_q = \frac{(x_q - \bar{x}_q)\sum_{r=1}^{N}(d_{qr} - d(X))(2x_r - 1) + d(x)}{\left(\sum_{r=1}^{N} x_r\right)\left(\sum_{r=1}^{N}\bar{x}_r\right) + (x_q - \bar{x}_q)\sum_{r=1}^{N}(2x_r - 1) - 1}$$

IF THE VALUE OF x_ℓ IS SWITCHED TO $\bar{x}_\ell$:

$\delta'_q = N/D$ WHERE

$$N = \delta_q[(Na - a) + (2x_q - 1)(2a - N) - 1] - \delta_\ell[(2x_q - 1)(2a - N) - 4x_\ell + 3] - 2(d_{q\ell} - d(X))(2x_\ell - 1)$$

$$D = a(N - a) + (2a - |N|)(2x_q + 2x_\ell - 2) - 2(2x_\ell - 1)(2x_q - 1) - 2$$

AND $\quad a = \sum_{r=1}^{N} x_r$

4. EXACT ALGORITHM

- DINKELBACH'S LEMMA (0-1 CASE)

CONSIDER A HYPERBOLIC 0-1 PROGRAM:

$$\max \frac{N(X)}{D(X)} \quad \text{WHERE} \quad D(X) > 0 \text{ FOR ALL } X.$$

LET $\lambda = \dfrac{N(X^0)}{D(X^0)}$. THEN X^0 IS OPTIMAL

IF AND ONLY IF $\min\limits_{X}(\lambda D(X) - N(X)) \geq 0$.

$\Rightarrow$ ITERATIVE SCHEME FOR BIPARTITIONING WITH MAXIMUM AVERAGE DISSIMILARITY

- ALGORITHM DALC (DIVISIVE AVERAGE-LINKAGE CLUSTERING)

a) INITIALIZATION

SELECT FEASIBLE SOLUTION X_0 (e.g. FROM TABU)

COMPUTE VALUE OF $\bar{d}(X_0)$

$X^* \leftarrow X_0$;

$\bar{d}^* \leftarrow \bar{d}(X_0)$;

- LET X^h BE A HEURISTIC SOLUTION OF THE FOLLOWING PROBLEM:

$$\min_{X} \ \bar{d}(X^0)\, D(X) - N(X) \qquad (Q)$$

IF $\bar{d}(X^0)\, D(X_1) - N(X_1) < 0$ REPLACE X_0 BY X_1 AND ITERATE

OTHERWISE SEEK AN OPTIMAL SOLUTION X_1 TO (Q) BY AN EXACT ALGORITHM (ENUMERATIVE OR CUTTING-PLANE ONE)

IF $\bar{d}(X_0)\, D(X_1) - N(X_1) \geq 0$ STOP ;

OTHERWISE ITERATE.

UPDATING OF CHANGE FORMULAE IS ALSO USED HERE

EXPERIMENTS

generated problems similar to those of the previous series of experiments are presented in Tables 9 and 10. It appears that: *(i)* HTABU gives slightly better values than HDALC; *(ii)* computing times of HTABU are smaller than those of HDALC. Therefore, a direct application of Tabu Search to the hyperbolic Problem (P) appears to be preferable to the Tabu Search heuristic solution of the sequence of quadratic 0-1 problems given by Dinkelbach's lemma (1967).

	BJ			HDALC			DALC		
Partition	D_m	$\overline{D}_m$	MD_m	D_m	$\overline{D}_m$	MD_m	D_m	$\overline{D}_m$	MD_m
P_1	776	776	776	793	793	793	793	793	793
P_2	763	794	763	773	753	683	773	753	683
P_3	[illegible]	767	[illegible]	780	775	683	780	775	683
P_4	646	750	646	677	750	646	677	750	646
P_5	542	752	542	542	752	542	542	732	542
P_6	532	732	532	532	732	532	532	732	532
P_7	465	717	465	465	717	465	465	717	465
P_8	436	704	436	436	704	436	436	704	436
P_9	415	709	415	415	709	415	415	709	415
P_{10}	330	702	330	333	699	333	333	699	333
P_{11}	278	692	278	300	691	277	300	691	277
P_{12}	277	679	277	277	679	277	277	679	277
tcpu (seconds)	0.01	0.02	0.01	0.10	0.11	0.11	0.31	0.32	0.32

SMALL PROBLEM: EXACT ALGO.

Table 5: Comparison of heuristics and exact algorithm (13 psychological tests of Harman)

	BJ			HDALC		
Partition	D_m	$\overline{D}_m$	MD_m	D_m	$\overline{D}_m$	MD_m
P_1	4062	4062	4062	4062	4062	4062
P_2	1963	3516	1963	2664	4345	2664
P_3	1785	3013	1785	2582	3752	880
P_4	1380	[illegible]	1380	2493	[illegible]	715
P_5	1313	3283	1313	2394	[illegible]	264
P_6	1186	3024	1186	2300	3156	264
P_7	[illegible]	3158	[illegible]	2256	3195	264
P_8	1088	3274	1088	2112	3025	264
P_9	1069	3144	1069	2022	2885	264
P_{10}	968	3061	968	1934	2733	264
P_{20}	706	2899	706	1506	2272	223
P_{30}	447	2853	447	719	2799	141
tcpu (seconds)	0.23	0.45	0.43	48.96	51.04	51.02

LARGE PROBLEM: TABU

Table 6: Heuristic solutions for Fisher's iris data (N=150)

AVERAGE DISSIMILARITIES													
		BJ			HDALC			QDALC			HQDALC		
N	M	D_m	$\overline{D}_m$	MD_m	D_m	$\overline{D}_m$	MD_m	D_m	$\overline{D}_m$	MD_m	D_m	$\overline{D}_m$	MD_m
100	2	170543	170958	170543	172412	173145	169610	172373	172763	168921	172711	172991	170798
	3	170287	170854	170287	171972	171997	164696	171999	172069	164472	172688	171329	161924
	4	169868	170743	169868	171591	172201	163344	171644	172401	163799	173036	171449	157567
	5	169532	170633	169532	171034	172421	161785	171183	172821	163473	172327	171195	151398
	10	167470	170407	167470	168408	172107	160279	168884	172324	159029	170146	170919	140221
	15	164597	170316	164597	168085	171470	154504	165696	171527	153028	169310	170793	124998
200	2	171659	175035	171489	173096	175208	171578	171659	175035	171489	17240[illegible]	171089	169176
	3	171458	174515	166151	171482	175350	169474	171458	174515	166151	173035	171597	163984
	4	171299	172602	154564	171221	173800	160512	171299	172602	154564	172231	172115	160359
	5	171087	172697	153683	171552	173340	157833	171087	172697	153683	171274	172390	158390
	10	170091	171686	143491	169725	172011	147302	170091	171686	143491	169777	171711	147035
	15	169642	171329	140815	169041	171308	145602	169642	171329	140815	169154	170805	131995
300	2	169477	169649	169477	172184	171367	167231	171167	172368	168407	174547	171268	167682
	3	169357	169594	169357	171417	172377	167109	171027	173107	168273	171600	172122	166424
	4	169259	169597	169259	170747	171754	163740	170730	172586	165254	171410	172389	159005
	5	169088	169541	169088	170349	172102	162619	170616	172466	158251	171976	171855	154072
	10	168270	169433	168270	169612	171573	153331	169722	171906	148969	171245	171881	138795
	15	167558	169386	167558	168444	171069	150221	168872	171054	136313	169801	170923	126190
400	2	169280	169372	169280	170861	170833	164545	170882	172334	170114	172828	172224	170489
	3	169227	169348	169227	171302	169744	159428	170645	172536	166749	171417	170785	161514
	4	169110	169324	169110	171275	170496	156932	170452	171111	158441	170413	170433	161028
	5	169024	169328	169024	171823	170480	153301	170225	170999	157266	170690	170296	153085
	10	168424	169264	168424	169844	170528	142275	169731	170970	140175	171375	170210	133555
	15	167787	169221	167787	168287	170375	139878	168982	170893	137292	169422	170171	121299
500	2	169183	169248	169183	170733	173179	170092	170628	171917	169928	171086	171038	169093
	3	169131	169235	169131	170567	171630	160770	170373	170182	159638	171890	170044	161369
	4	169036	169206	169036	170375	171235	156738	170233	170816	156603	169967	169353	152531
	5	168927	169198	168927	170731	171272	149648	170115	171044	151156	171144	169624	148821
	10	168485	169170	168485	169803	170748	138304	169532	170244	128375	169929	169750	135575
	15	167941	169140	167941	169091	170404	137526	169688	170359	125768	169881	169522	126278

AVERAGE TIME												
	BJ			HDALC			QDALC			HQDALC		
N	D_m	$\overline{D}_m$	MD_m	D_m	$\overline{D}_m$	MD_m	D_m	$\overline{D}_m$	MD_m	D_m	$\overline{D}_m$	MD_m
100	0.54	0.65	0.62	10.37	10.92	10.91	51.23	51.76	51.75	46.05	46.72	46.70
200	1.06	1.53	1.49	47.75	54.97	54.97	144.36	149.90	149.95	138.73	144.14	143.59
300	2.14	3.37	3.22	67.21	85.94	85.84	230.48	256.61	256.50	211.89	229.12	229.08
400	3.60	6.37	5.88	130.12	183.23	183.17	342.13	384.36	384.24	326.01	364.84	364.78
500	6.17	9.38	8.42	147.03	242.90	241.88	602.21	679.18	678.05	585.09	661.33	660.19

Table 9: Results for large problems

Image Representation, Generalized Clustering, and Search in Proximity Graphs and Pathfinder Networks

Donald W. Dearholt
Department of Computer Science
Mississippi State University
MS 39762

e-mail address: dearholt@cs.msstate.edu
Phone: 601-325-2756

Abstract

Pathfinder associative networks (PFNETs) were originated to model human semantic memory, and have proven particularly effective with associative aspects of the organization of knowledge. Theoretical connections have been established with graph theory, path algebras, proximity graphs, computational geometry, clustering, and search procedures. The PFNET paradigm is being used in human-computer interface systems, and in the organization of data in a database for an experimental robotic vision system. Characteristics of PFNETs include (1) the preservation of minimum-distance paths between entities, (2) the clustering of similar entities through the edge structure, (3) the consequent support of higher levels of abstraction, and (4) the capability of generating proximity graphs, such as the relative neighborhood graph and the (open lune) gabriel graph. Monotonic search networks (MSNETs), closely related to PFNETs, provide for search in which no backtracking is ever necessary in domains having (objective) distance measures, and also support the clustering features of PFNETs. These studies have motivated investigations of cluster learning and conceptual clustering from the perspective of primitive transformations, with representation of the clusterings by means of PFNETs. Models of cluster learning in which the clusters represent either ordered or unordered sets, and which may or may not overlap, are being considered. The sequential application of transformations on clusters to a graph-generation algorithm is being considered as a learning paradigm, and this perspective appears to support a constructive view of clustering. Co-occurrence of entities is an essential component of the process.

PROPERTIES AND APPLICATIONS OF PATHFINDER-BASED ASSOCIATIVE NETWORKS

Don Dearholt

DEPARTMENT OF COMPUTER SCIENCE
MISSISSIPPI STATE UNIVERSITY

Developed under the sponsorship of:

AIR FORCE HUMAN RESOURCES LABORATORY

INSTRUMENTATION DIRECTORATE,
WHITE SANDS MISSILE RANGE

COMPUTING RESEARCH LABORATORY, NMSU

INTERNATIONAL BUSINESS MACHINES, INC.

NATIONAL SCIENCE FOUNDATION

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

OFFICE OF NAVAL RESEARCH

PHILOSOPHICAL STANCE: BETTER MODELING OF HUMAN INTELLIGENCE WILL LEAD TO BETTER AI

THE NETWORKS WE ARE STUDYING:

DESCRIBE, SUMMARIZE, AND DISPLAY DATA

SUGGEST A PSYCHOLOGICAL MODEL ABOUT MENTAL REPRESENTATIONS

COMPLEMENT MDS AND CLUSTER ANALYSIS

PROVIDE A PARADIGM FOR:

- KNOWLEDGE REPRESENTATION
- MODELS OF CLASSIFICATION
- ORGANIZATION OF DATABASE SYSTEMS
- SPREADING ACTIVATION (SEARCH)

OUTLINE

I. MOTIVATION, PERSPECTIVE, AND OBJECTIVES

II. PATHFINDER NETWORKS

A. DEFINITIONS AND PROPERTIES

B. APPLICATIONS

III. CLUSTER LEARNING AND DYNAMIC SYSTEMS

A. MOTIVATIONS AND APPROACH

B. DEFINITIONS AND PROPERTIES

C. APPLICATIONS

RESEARCH OBJECTIVES

I. THEORETICAL

DEVELOP AND TEST METRICS

RELATIONSHIPS:

- GRAPH THEORY
- PATH ALGEBRAS
- PROXIMITY GRAPHS (RNG, GG, DTG)

LEVELS OF ABSTRACTION

II. EMPIRICAL

SEMANTIC MEMORY

CLASSIFICATION MODELS

PROPOSITIONAL ANALYSIS

KNOWLEDGE EXTRACTION FROM EXPERTS

III. APPLICATION DOMAINS

ORGANIZATION OF CONCEPTS

- INTERFACES--INFORMATION RETRIEVAL, HELP SYSTEMS
- DATABASE ORGANIZATION

PERCEPTION--OUTLINES OF OBJECTS

THE BIGGEST CHALLENGE

FOR AI AND COGNITIVE MODELING:

TO DESIGN A SYSTEM WHICH DOES MANY THINGS WELL, ALTHOUGH EACH ALGORITHM MIGHT NOT BE OPTIMAL

- ASSOCIATIONAL ORGANIZATION
- CLUSTERING
- SEVERAL LEVELS OF ABSTRACTION
- CLASSIFICATION
- SEARCH
- DESCRIPTION OF DECISIONS

DEFINITION

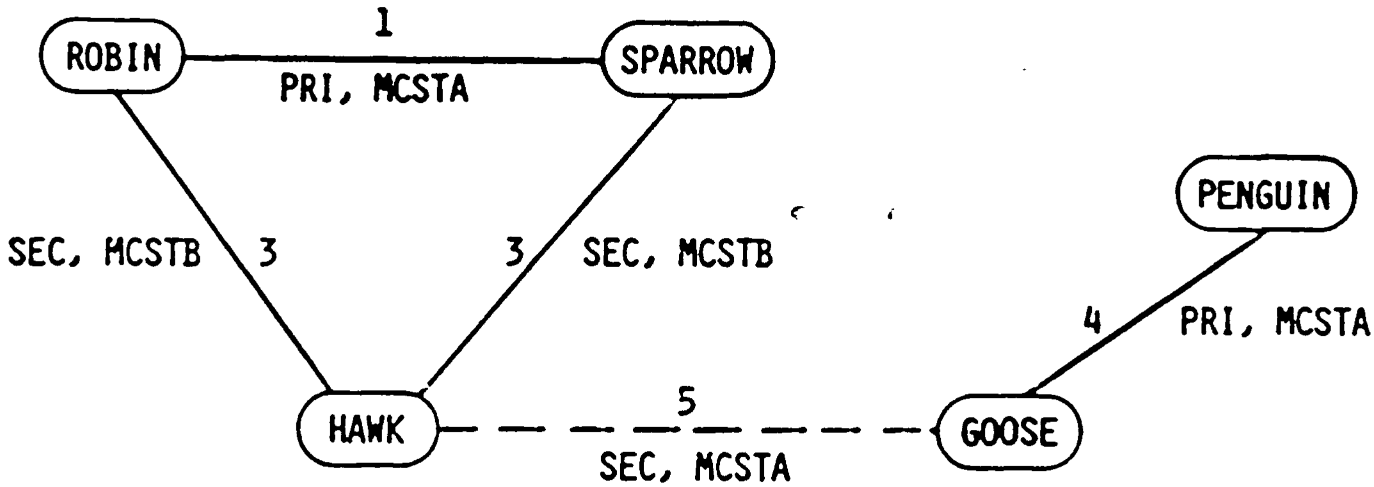
A PATHFINDER NETWORK (PFNET) IS A GRAPH BASED ON PAIRWISE ESTIMATES OR MEASURES OF DISTANCES BETWEEN ENTITIES.

EACH ENTITY CORRESPONDS TO A NODE.

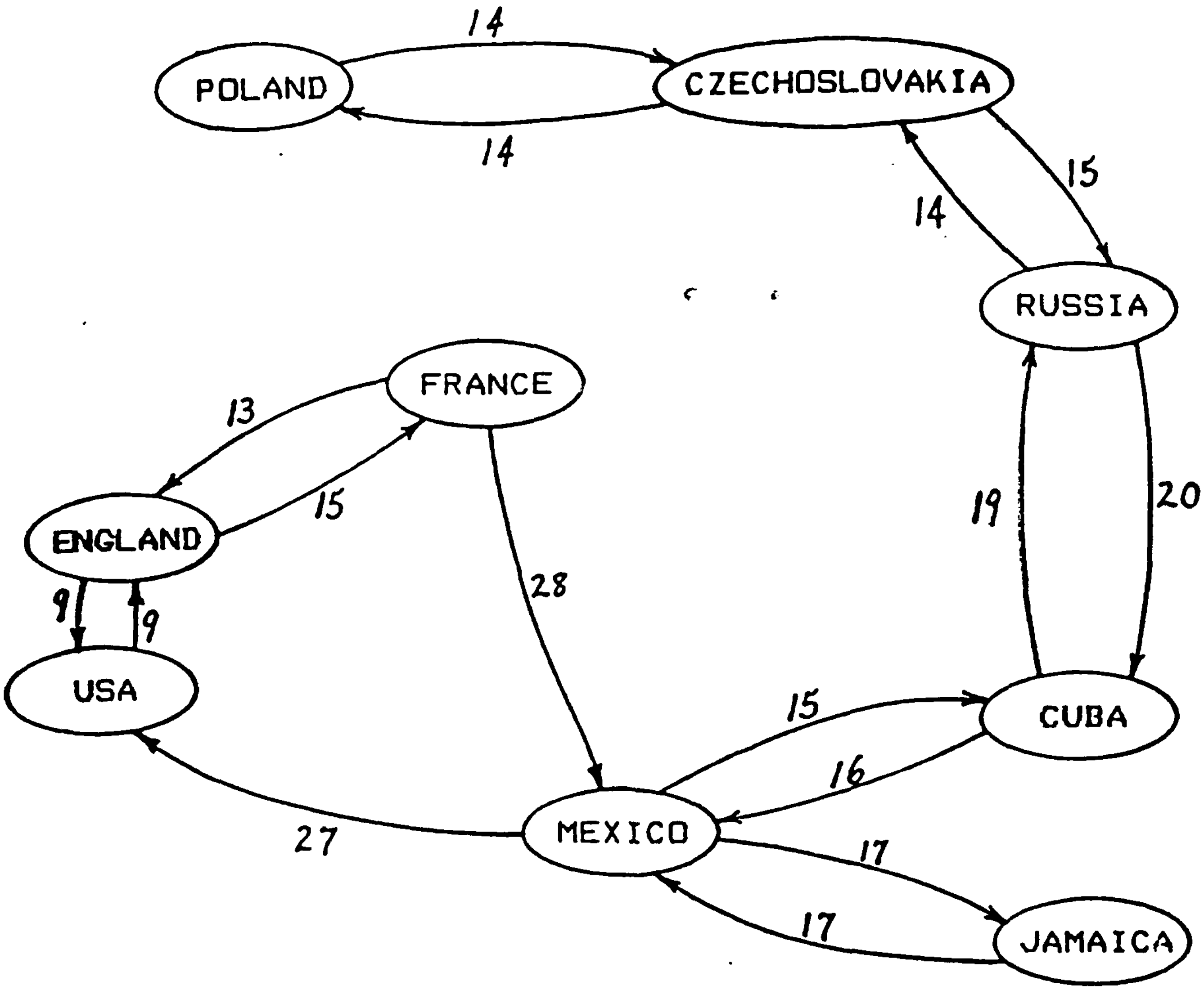
EACH PAIR OF NODES IN A PFNET IS CONNECTED DIRECTLY BY AN EDGE WHOSE WEIGHT IS THE DISTANCE BETWEEN THE TWO ENTITIES,

UNLESS THERE IS A SHORTER ALTERNATIVE PATH.

EXAMPLE OF A LABELED PFNET



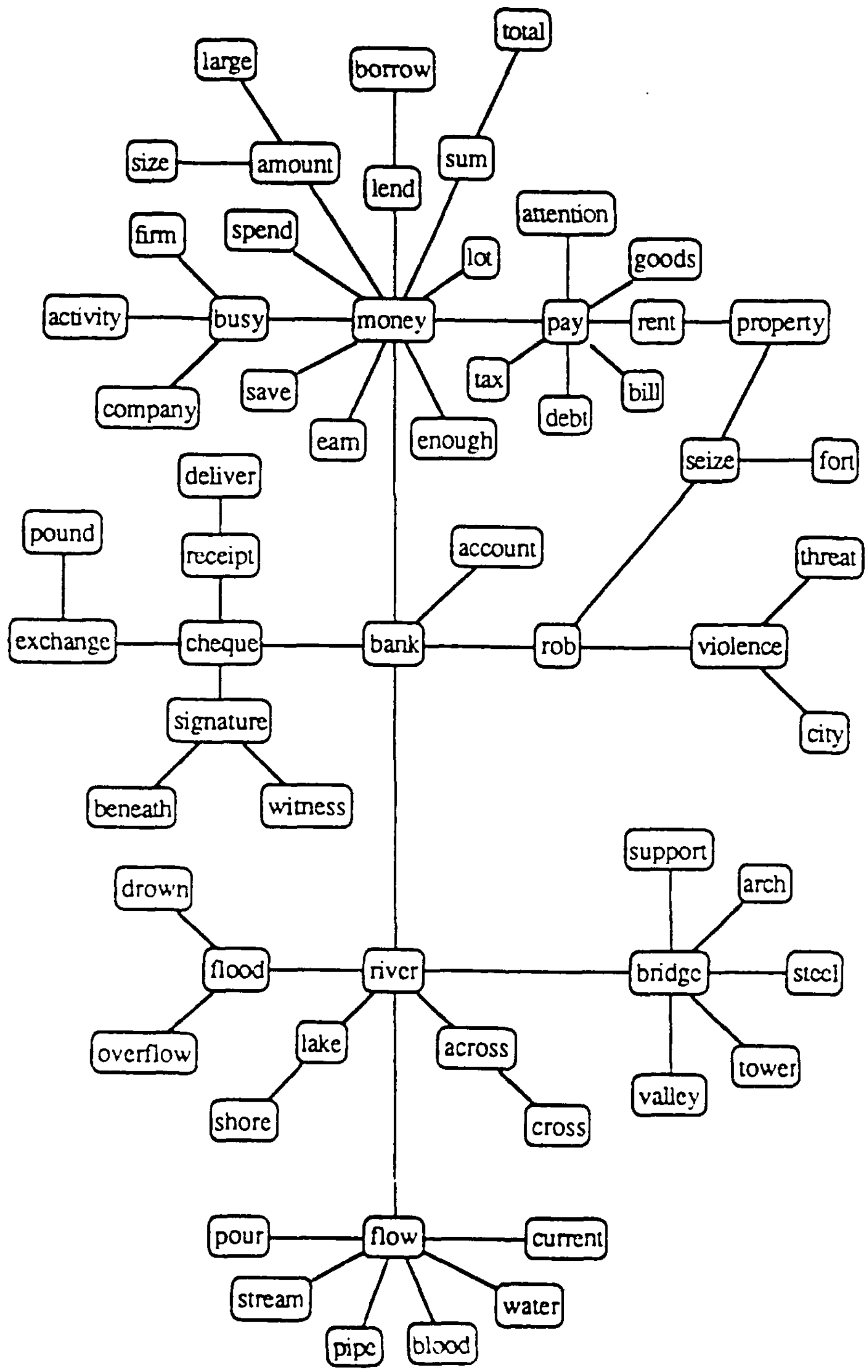
DIRECTED PFNET FOR NINE COUNTRIES



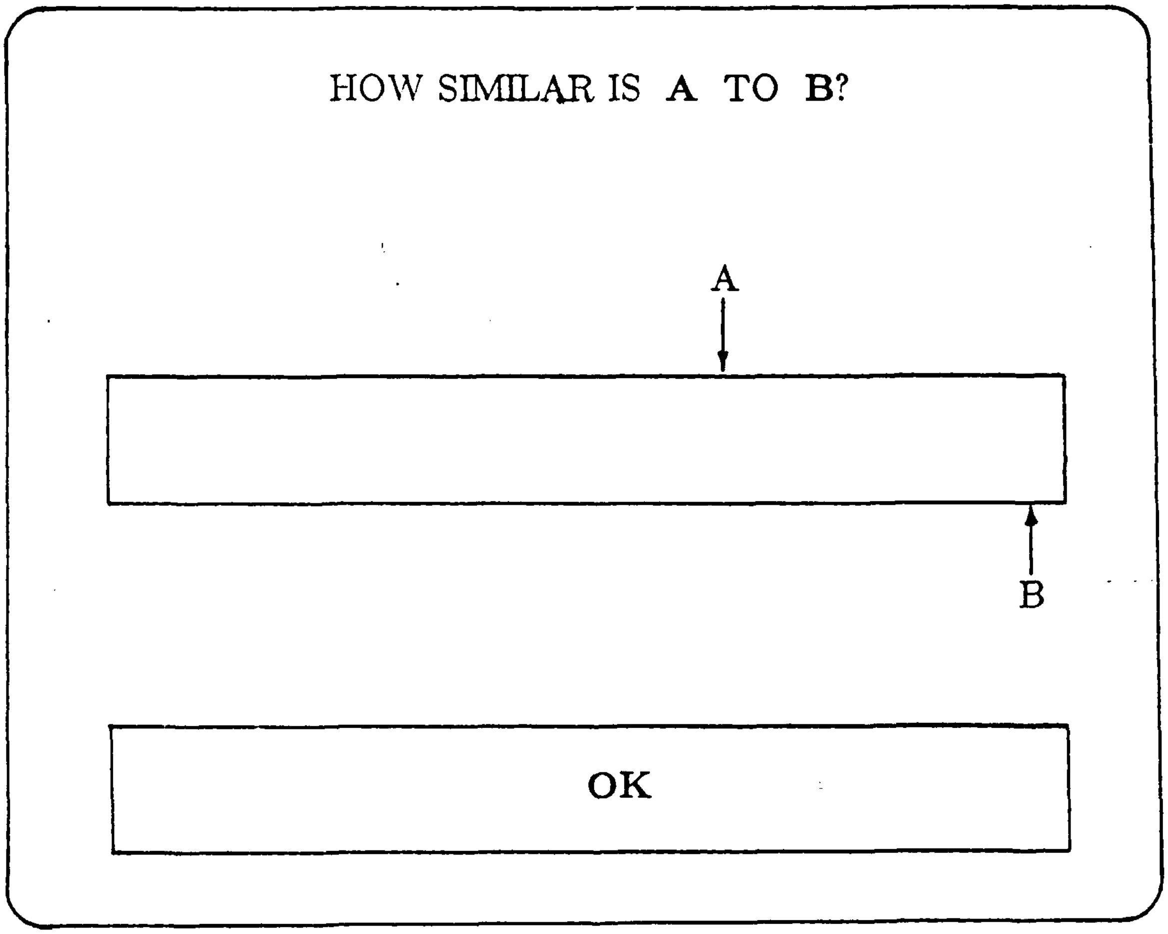
R-METRIC IS INFINITY

Q-PARAMETER IS EIGHT

LDOCE: AMBIGUITY RESOLUTION



TOUCHSCREEN DISPLAY FOR EMPIRICAL DATA



$$DISTANCE + SIMILARITY = K$$

THE PARAMETERS OF A PFNET

R-METRIC:

RULE FOR FINDING THE LENGTH OF A PATH WITH K EDGES

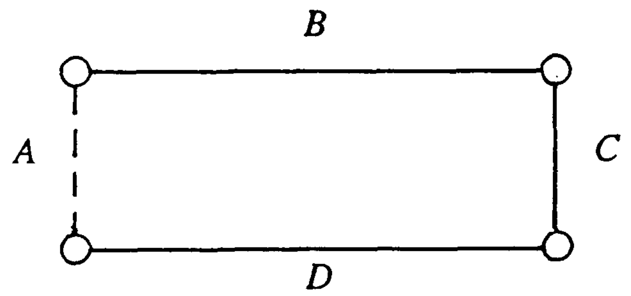
$$L(P) = \left[\sum_{I=1}^{K} W_I^{R} \right]^{1/R}$$

R	PATH LENGTH	DATA SCALE
1	SUM OF WEIGHTS	RATIO
2	EUCLIDEAN	RATIO
•		
•		
∞	MAXIMUM WEIGHT	RATIO, ORDINAL

THE PARAMETERS OF A PFNET

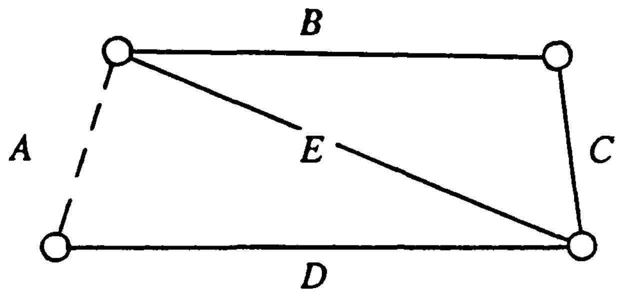
Q-PARAMETER:

"DIMENSION" OF GENERALIZED TRIANGLE INEQUALITIES SATISFIED



$$A \leq \left[B^R + C^R + D^R\right]^{1/R}$$

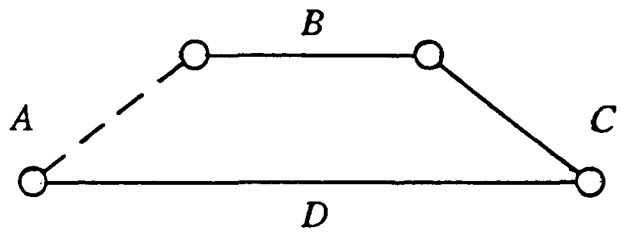
THE TRIANGLE INEQUALITY



$E \leq B + C$

$A \leq E + D \leq B + C + D$

THE GENERALIZED TRIANGLE INEQUALITY



$A \leq [B^R + C^R + D^R]^{1/R}$

PURPOSE: TO PRESERVE MINIMAL-DISTANCE PATHS

THEORETICAL RESULTS

FOR A GIVEN DISTANCE MATRIX,

PFNET(R, Q):

IS UNIQUE,

PRESERVES GEODETIC DISTANCES,

LINKS NEAREST NEIGHBORS, AND

CONTAINS THE SAME INFORMATION AS THE MINIMUM METHOD OF HIERARCHICAL CLUSTERING

PFNET($R=\infty$, $Q=N-1$) IS THE UNION OF ALL MINTREES

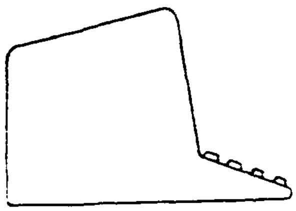
PFNET(R_2, Q) IS A SPANNING SUBGRAPH OF PFNET(R_1, Q) IFF $R_1 \leq R_2$

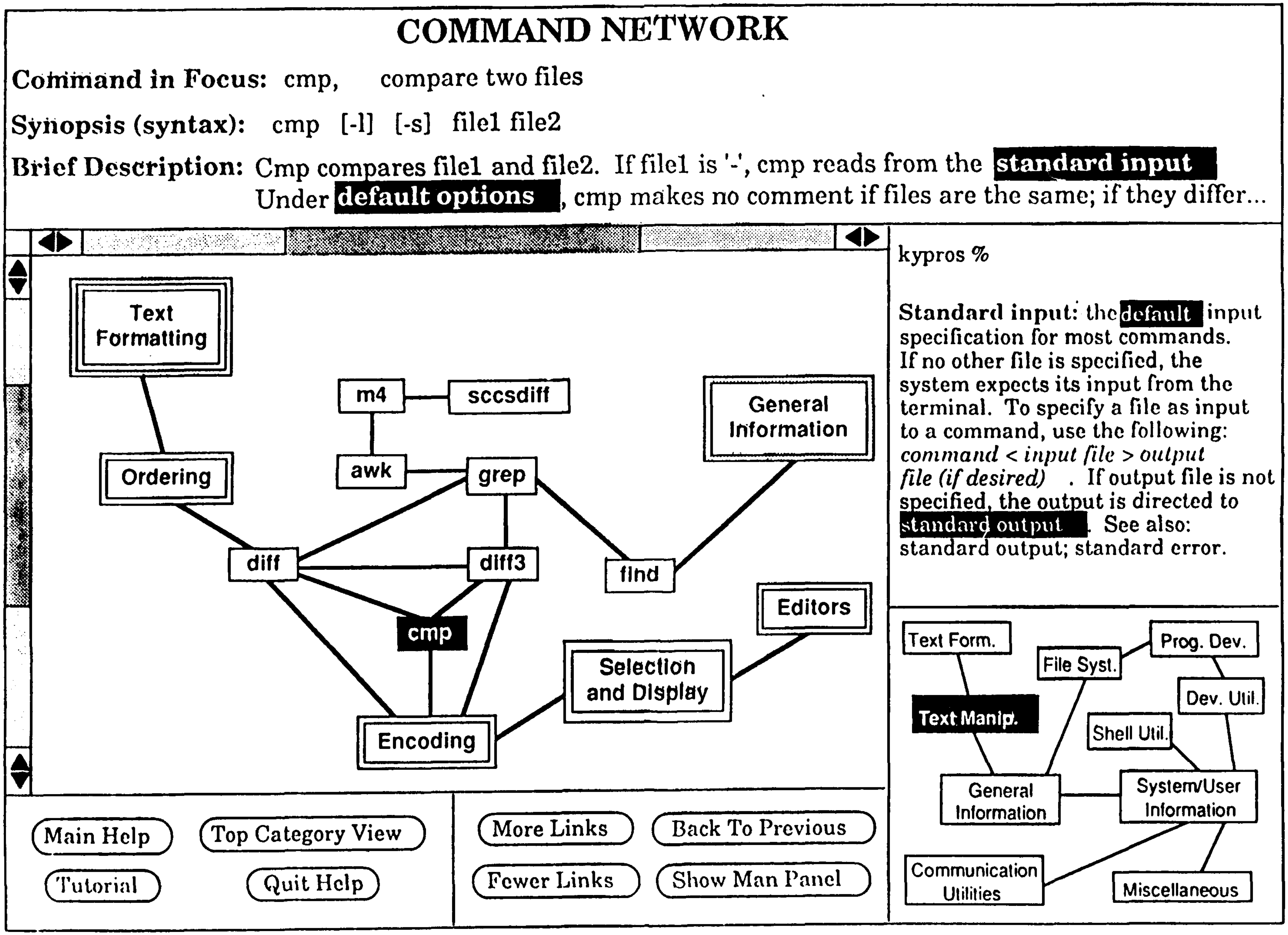
PFNET(R, Q_2) IS A SPANNING SUBGRAPH OF PFNET(R, Q_1) IFF $Q_1 \leq Q_2$

MONOTONIC TRANSFORMATIONS PRESERVE STRUCTURE FOR ALL PFNET($R=\infty$, Q)

MULTIPLICATIVE TRANSFORMATIONS PRESERVE STRUCTURE FOR ALL PFNET(R, Q)

T H E I
U N
C O M P U T E R
A E
N R
F

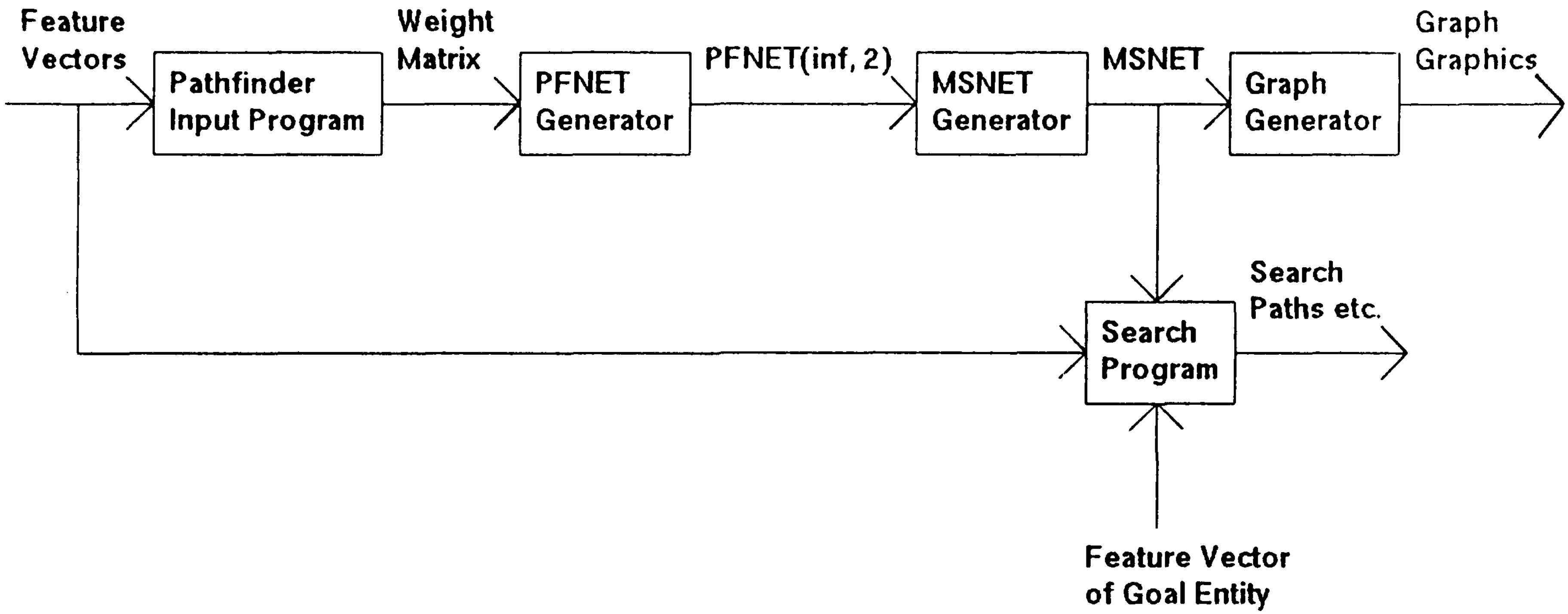
A
C

E


COMMAND NETWORK
Command in Focus: cmp, compare two files
Synopsis (syntax): cmp [-l] [-s] file1 file2
Brief Description: Cmp compares file1 and file2. If file1 is '-', cmp reads from the standard input
Under default options, cmp makes no comment if files are the same; if they differ...
Text Formatting
m4
sccsdiff
General Information
Ordering
awk
grep
diff
diff3
find
Editors
cmp
Selection and Display
Encoding
Main Help
Top Category View
Tutorial
Quit Help
More Links
Back To Previous
Fewer Links
Show Man Panel
kypros %
Standard input: the default input specification for most commands. If no other file is specified, the system expects its input from the terminal. To specify a file as input to a command, use the following: command < input file > output file (if desired) . If output file is not specified, the output is directed to standard output . See also: standard output; standard error.
Text Form.
File Syst.
Prog. Dev.
Dev. Util.
Text Manip.
Shell Util.
General Information
System/User Information
Communication Utilities
Miscellaneous

COMPUTER VISION

GOAL: SCAN THE ENVIRONMENT AND MAKE DECISIONS WITHOUT HUMAN INTERACTION

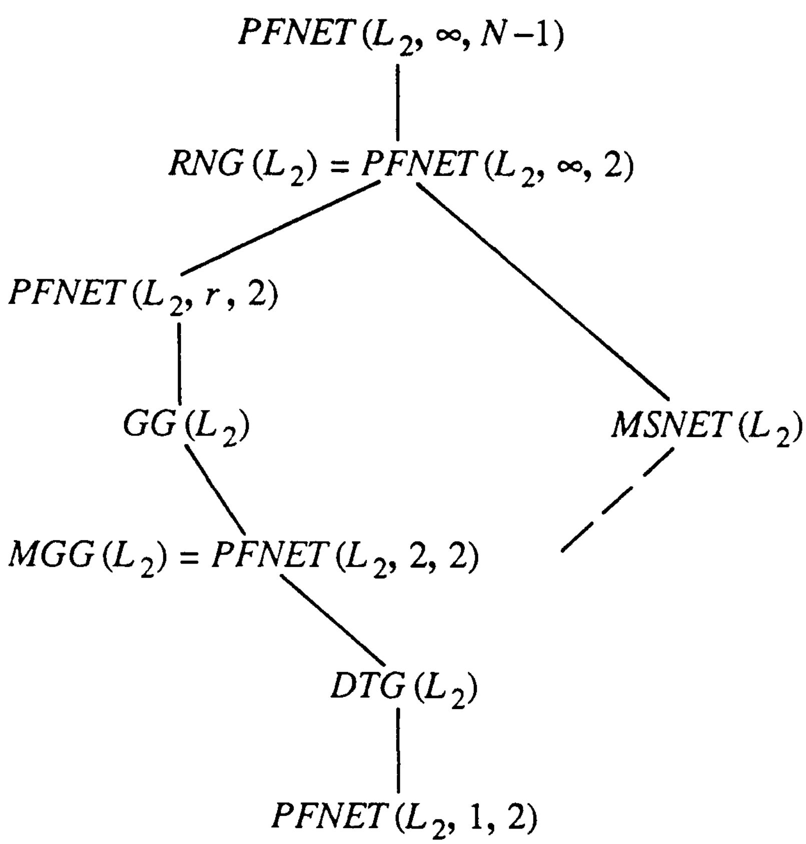
REQUIRES: KNOWLEDGE REPRESENTATION
CLASSIFICATION
ABILITY TO *DESCRIBE* SCENE
RECONSTRUCT SCENE
ENHANCE SCENE
MODIFY SCENE



SYSTEM ORGANIZATION FOR THE DATABASE

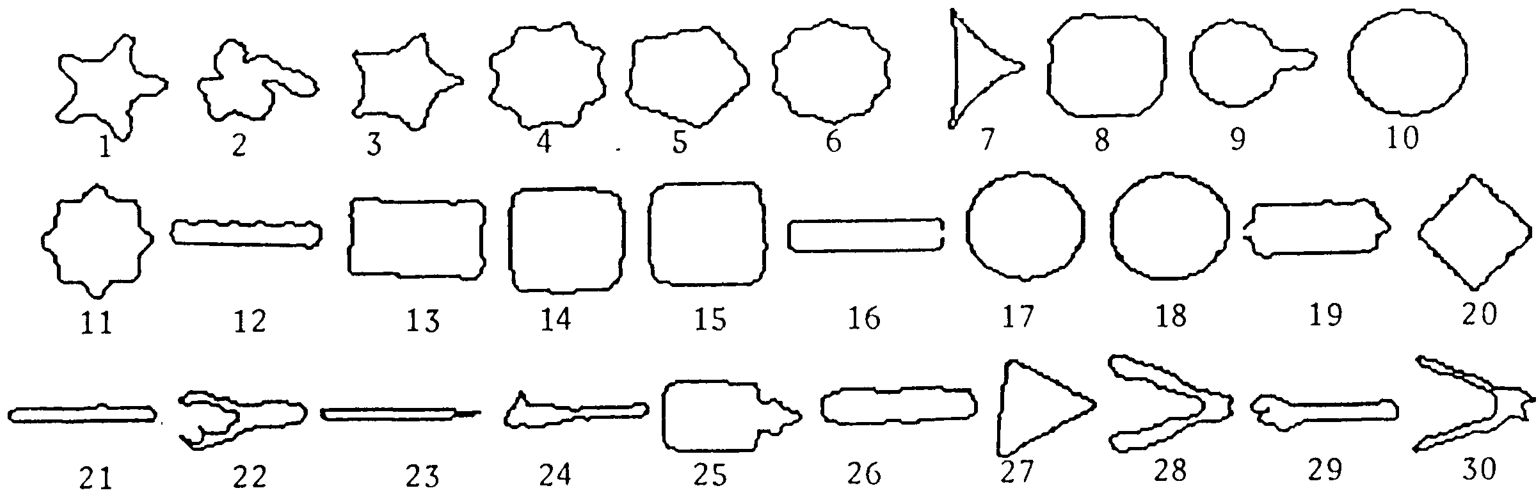
A HIERARCHY OF

EMPTY NEIGHBORHOOD GRAPHS

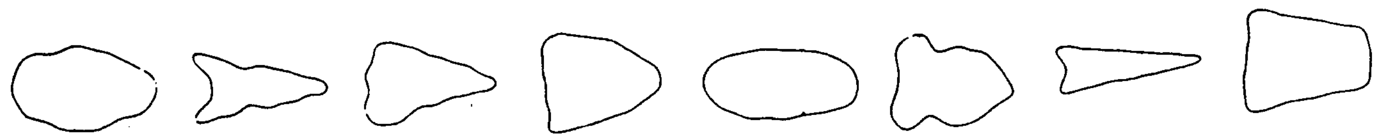


EACH GRAPH IS A SPANNING SUBGRAPH

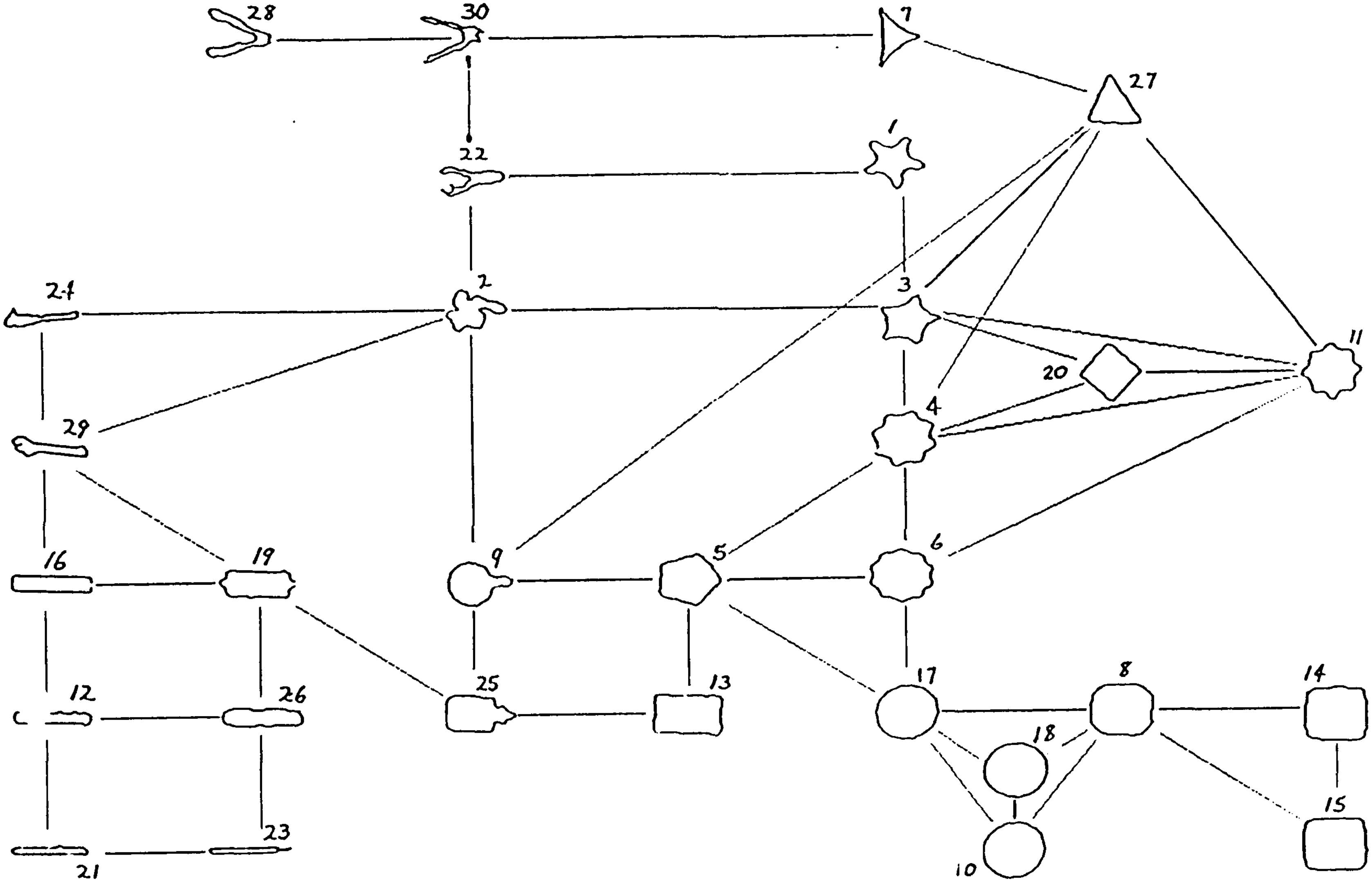
OF THE GRAPH BELOW IT



DATABASE OBJECTS



EIGHT ARCHEOLOGICAL TEST OBJECTS



MSNET ORGANIZATION OF THE DATA

MOTIVATIONS FOR STUDY OF CLUSTERINGS

LEARNING VIA CLUSTERINGS

MODEL DYNAMIC SYSTEMS

UTILIZE PATHFINDER NETWORKS
- ASSOCIATIVITY
- CLUSTERING
- LEVELS OF ABSTRACTION

CONCEPT/CLUSTER LEARNING MODEL SHOULD SUPPORT:

SETS OF CLUSTERS AS THE DOMAIN

PROGRESSION THRU CLUSTERINGS AS LEARNING OCCURS

A DISTANCE MEASURE BETWEEN CLUSTERINGS

INTERPRETATION WITHIN THE PATHFINDER PARADIGM

MODEL FOR DYNAMIC SYSTEMS SHOULD SUPPORT:

SETS OF CLUSTERS AS THE DOMAIN

PROGRESSION THRU CLUSTERINGS AS SYSTEM CHANGES

A DISTANCE MEASURE BETWEEN CLUSTERINGS

PATHFINDER OR PROXIMITY GRAPH PARADIGM

ADVANTAGES

PSYCHOLOGICAL FOUNDATIONS

REPETITION OR REHEARSAL CAN BE MODELED

REPRESENTATIONS AS ASSOCIATIVE GRAPHS

KNOWN GRAPH-THEORETIC PROPERTIES

VERIDICAL GRAPHICAL DISPLAY

ASSUMPTIONS

NOTATION DENOTES CLUSTERS VIA ( )s

CO-OCCURRENCES DERIVED FROM ( )s NOTATION

FEATURES OF ENTITIES WON'T BE CONSIDERED

REPETITION INCREMENTS CO-OCCURRENCE

CO-OCC + DISSIM = CONSTANT

TWO MODELS:

1. SEQUENTIAL LEARNING PARADIGM
2. CONCURRENT (SNAPSHOT) PARADIGM

METRIC AXIOMS

REFLEXIVITY

SYMMETRY

TRIANGLE INEQUALITY--INVALID FOR CO-OCCURRENCES

PFNs IMPOSE THE GENERALIZED TRIANGLE INEQUALITY

COMPONENTS OF THE MODEL

DOMAIN $D = \{A, B, C, D, \ldots\}$

SAMPLE CLUSTERING $Ci = \{(A, B, C), D, E\}$

PRIMITIVE OPERATIONS

- ADD ENTITY TO D
- REMOVE ENTITY FROM D
- MERGE TWO CLUSTERS INTO ONE CLUSTER
- SPLIT ONE CLUSTER INTO TWO CLUSTERS

WHERE AN ENTITY IS A CLUSTER

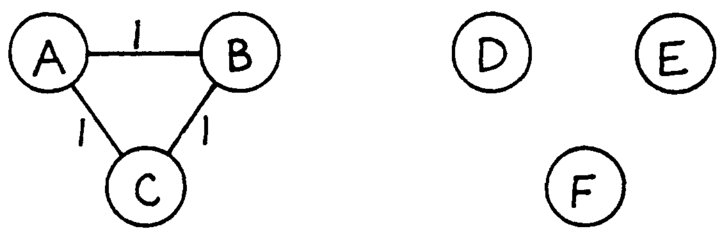
EXAMPLE: THE LEARNING PARADIGM

C0 = {A, B, C, D, E, F}

$$S0 = \begin{matrix} - & 0 & 0 & 0 & 0 & 0 \\ 0 & - & 0 & 0 & 0 & 0 \\ 0 & 0 & - & 0 & 0 & 0 \\ 0 & 0 & 0 & - & 0 & 0 \\ 0 & 0 & 0 & 0 & - & 0 \\ 0 & 0 & 0 & 0 & 0 & - \end{matrix}$$

C1 = {(A, B, C), D, E, F}

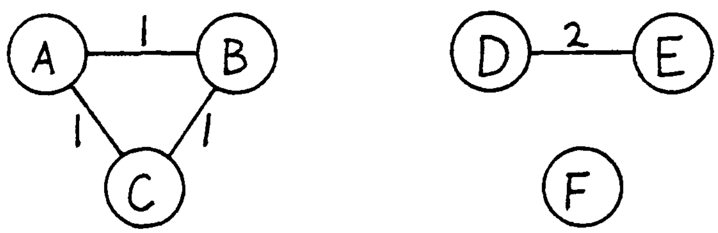
$$S1 = \begin{matrix} - & 1 & 1 & 0 & 0 & 0 \\ 1 & - & 1 & 0 & 0 & 0 \\ 1 & 1 & - & 0 & 0 & 0 \\ 0 & 0 & 0 & - & 0 & 0 \\ 0 & 0 & 0 & 0 & - & 0 \\ 0 & 0 & 0 & 0 & 0 & - \end{matrix}$$



EXAMPLE CONTINUED:

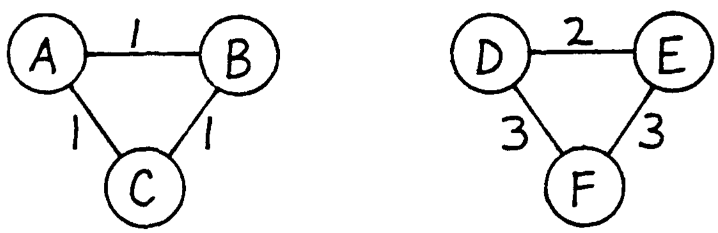
C2 = {(A, B, C)2, (D, E), F}

$$S2 = \begin{array}{cccccc} - & 2 & 2 & 0 & 0 & 0 \\ 2 & - & 2 & 0 & 0 & 0 \\ 2 & 2 & - & 0 & 0 & 0 \\ 0 & 0 & 0 & - & 1 & 0 \\ 0 & 0 & 0 & 1 & - & 0 \\ 0 & 0 & 0 & 0 & 0 & - \end{array}$$



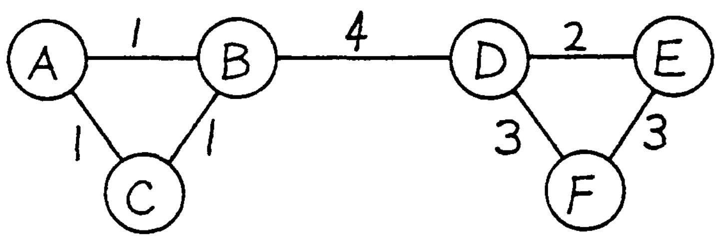
C3 = {(A, B, C)3, ((D, E), F)}

$$S3 = \begin{array}{cccccc} - & 3 & 3 & 0 & 0 & 0 \\ 3 & - & 3 & 0 & 0 & 0 \\ 3 & 3 & - & 0 & 0 & 0 \\ 0 & 0 & 0 & - & 2 & 1 \\ 0 & 0 & 0 & 2 & - & 1 \\ 0 & 0 & 0 & 1 & 1 & - \end{array}$$



EXAMPLE CONTINUED:

C4 = {(A, B, C)4, ((D, E), F)2, (B, D)}

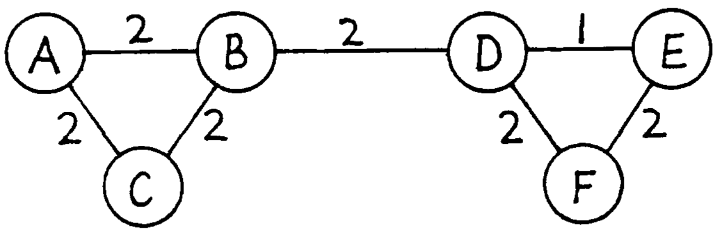


$$S4 = \begin{matrix} - & 4 & 4 & 0 & 0 & 0 \\ 4 & - & 4 & 1 & 0 & 0 \\ 4 & 4 & - & 0 & 0 & 0 \\ 0 & 1 & 0 & - & 3 & 2 \\ 0 & 0 & 0 & 3 & - & 0 \\ 0 & 0 & 0 & 2 & 0 & - \end{matrix}$$

EXAMPLE: THE CONCURRENT PARADIGM

C = {(A, B, C), ((D, E), F), (B, D)}

$$
S0 = \begin{matrix}
- & 1 & 1 & 0 & 0 & 0 \\
1 & - & 1 & 1 & 0 & 0 \\
1 & 1 & - & 0 & 0 & 0 \\
0 & 1 & 0 & - & 2 & 1 \\
0 & 0 & 0 & 2 & - & 0 \\
0 & 0 & 0 & 1 & 0 & -
\end{matrix}
$$



APPLICATIONS

METRIC FOR DIFFERENCES IN EXPERTISE

IDENTIFICATION OF ERRORS

CONSENSUS

AUTOMATIC SYSTEM ADAPTATION WITH LEARNING
- HUMAN-COMPUTER INTERFACE
- ROBOTICS VISION DATABASE

TRANSFORMATIONS, DISTANCE, AND DISTANCE GRAPHS

Gary Chartrand
Western Michigan University

ABSTRACT

Several transformations are described – between graphs and between subgraphs in a graph. Each of these transformations gives rise to a distance (between graphs or between subgraphs in a graph). The relations between collections of graphs or between the subgraphs of a specified size within a (connected) graph can be described by graphs themselves, called distance graphs. In addition to describing these concepts, another distance between induced subgraphs of a specified order and the corresponding distance graphs are also discussed.

Introduction

The distance between two vertices in a connected graph is the length of the shortest path connecting the vertices. Distance is one of the most fundamental concepts in graph theory. Algorithms for determining distance in graphs are well known while applications involving distance in graphs are varied and numerous. Indeed, so much work has been done on this subject that Buckley and Harary wrote a book in 1990 entirely devoted to distance in graphs. Distance in graphs has been generalized in several ways, most notably perhaps to Steiner distance.

The combination of distance and graphs occurs in many other ways. One of these concerns distance between graphs. It would be more accurate to speak of distance between certain pairs of graphs because in many instances distance is defined between graphs having some specified properties. In all such distances, the distance between two graphs is 0 if and only if they are isomorphic. Hence, the distance between two graphs is a measure of the structural difference between the graphs.

Transformations

Some distances between graphs involve the idea of transformations. Let G and H be two (p, q) graphs. We say that G can be *transformed into H by an edge rotation* if G contains distinct vertices u, v, and w such that $uv \in E(G)$, $uw \notin E(G)$, and $H \cong G - uv + uw$. More generally, we say that G can be *r-transformed* into H if there exists a sequence $G \cong G_0$, $G_1, \ldots, G_n \cong H$ $(n \geq 0)$ of graphs such that G_i can be transformed into G_{i+1} by an edge rotation for $i = 0, 1, \ldots, n-1$.

An edge slide is a restricted version of an edge rotation. A graph G can be *transformed into a graph H by an edge slide* if G contains distinct vertices u, v, and w such that $uv \in E(G)$, $vw \in E(G)$, $uw \notin E(G)$ and $H \cong G - uv + uw$. If a graph H is isomorphic to a graph G or H can be obtained from a graph G by a sequence of edge slides, we say that G can be *s-transformed* into H. For example, the graph H of Figure 1 can be obtained from the graph G by an edge rotation, but H *cannot* be obtained from G by an edge slide. On the other hand, the graph H' can be obtained from G' by an edge slide (as well as by an edge rotation).

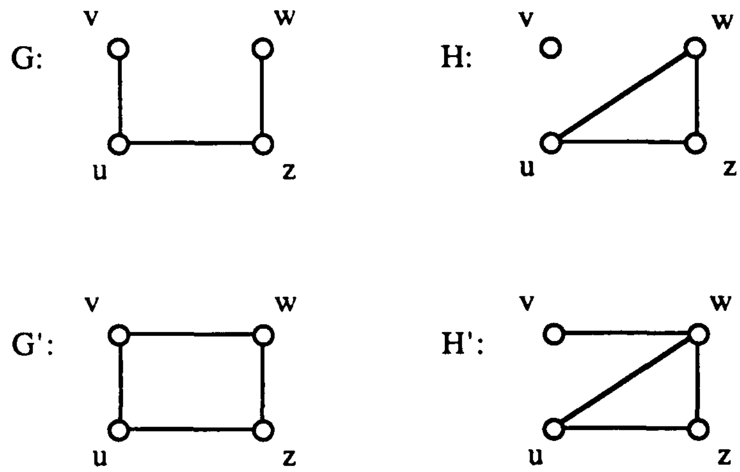


Figure 1

It was shown by Chartrand, Saba, and Zou that every (p, q) graph G can be r-transformed into any other (p, q) graph H. It was also shown by Johnson that s-transformation preserves connectedness. Further, a graph G can be s-transformed into a graph H if and only if G and H have the same number of components and corresponding components of G and H have the same order and same size.

Metrics Based on Transformations

Associated with these transformations are two metrics defined on graphs. Let G and H be two graphs having the same order and same size. The *edge rotation distance* or, more simply, the *r-distance* $d_r(G, H)$ between G and H is the smallest nonnegative integer n for which there exists a sequence $G_0, G_1, ..., G_n$ of graphs such that $G \cong G_0$, $H \cong G_n$, and G_i can be obtained from G_{i-1} by an edge rotation for $i = 1, 2, ..., n$. For example, the edge rotation distance between graphs G and H shown in Figure 2 is $d_r(G, H) = 3$.

Figure 2

The following properties of edge rotation distance were established by Chartrand, Saba, and Zou.

Proposition 1 If G and H are two graphs having the same order and same size, then $d_r(G, H) = d_r(\overline{G}, \overline{H})$.

It was shown that every nonnegative integer is the r-distance between some pair of graphs.

Proposition 2 For every nonnegative integer n, there exist graphs G and H such that $d_r(G, H) = n$.

Prior to presenting an upper bound for the r-distance between two graphs, we introduce another concept. For nonempty graphs G_1 and G_2, a *greatest common subgraph* of G_1 and G_2 is defined as any graph G of maximum size without isolated vertices that is (isomorphic to) a subgraph of both G_1 and G_2.

Proposition 3 Let G and H be two (p, q) graphs with $q \geq 1$, and let s be the size of a greatest common subgraph of G and H. Then $d_r(G, H) \leq 2(q - s)$. Moreover, this bound is sharp.

Another distance between graphs is associated with edge slide and was discussed by Johnson and by Benadé, Goddard, McKee, and Winter. Let G be a graph with components G_i, $1 \leq i \leq k$, and H a graph with components H_i, $1 \leq i \leq k$, such that G_i and H_i have the same order and same size. We define the *edge slide distance* or, simply, the *s-distance* $d_s(G, H)$ between G and H as the smallest nonnegative integer n for which there exists a sequence $G \cong G_0, G_1, ..., G_n \cong H$ of graphs such that, for $i = 1, 2, ..., n$, G_i can be obtained from G_{i-1} by an edge slide. If G and H are the graphs presented in Figure 3, then the edge slide distance between G and H is $d_s(G, H) = 2$.

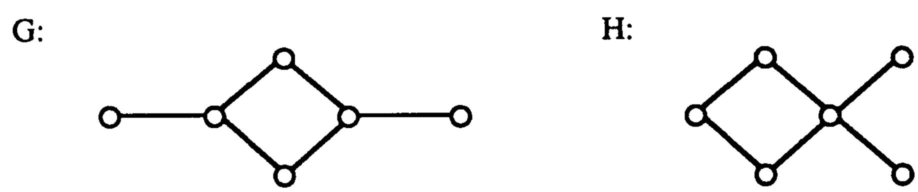


Figure 3

Note that $d_r(G, H) = 1$ for the graphs G and H of Figure 3. It is straightforward to show that $d_r(G, H) \leq d_s(G, H)$ for every pair G, H of connected graphs having the same order and same size. The following result is perhaps less obvious.

Proposition 4 For every pair m, n of positive integers with $m \leq n$, there exist graphs G and H such that $d_r(G, H) = m$ and $d_s(G, H) = n$.

Distance Graphs

Let S be a set of (nonisomorphic) (p, q) graphs. Then we define the *edge rotation distance graph* $\mathcal{D}_r(S)$ of S as the graph with the vertex set S such that two vertices G and H of $\mathcal{D}_r(S)$ are adjacent if and only if $d_r(G, H) = 1$. A graph F is *an edge rotation distance graph* if $F \cong D_r(S)$ for some set S of graphs.

Let S′ be a set of (nonisomorphic) graphs having the same number of components, labeled in such a way that the ith components of all graphs have the same order and same size. Then we define the *edge slide distance graph* $\mathcal{D}_s(S')$ of S′ analogously.

It was shown by Chartrand, Goddard, Henning, Lesniak, Swart, and Wall that every graph is an edge slide distance graph and it was conjectured that all graphs are edge rotation distance graphs. A number of classes of graphs are known to be edge rotation distance graphs. The next two results are due to Chartrand, Goddard, Henning, Lesniak, Swart, and Wall.

Proposition 5 Complete graphs, cycles and trees are edge rotation distance graphs.

Proposition 6 Every line graph is an edge rotation distance graph.

Proposition 7 (Faudree, Schelp, Lesniak, Gyárfás, and Lehel) The complete bipartite graphs $K_{3,3}$ and $K_{2,p}$ $(p \geq 1)$ are edge rotation distance graphs.

Proposition 8 (Jarrett) For every pair m, n of positive integers, the graph $K_{m,n}$ is an edge rotation distance graph.

F–Transformations

Let G and H be two (p, q) graphs, both containing a subgraph isomorphic to a given graph F of order at least 2. We say that G can be transformed into H by an *F-rotation* (or simply, G can be *F-rotated* into H) if there exist distinct vertices u, v, and w of G and a subgraph F′ of G isomorphic to F such that $u \notin V(F')$, $\{v, w\} \subseteq V(F')$, $uv \in E(G)$, $uw \notin E(G)$, and $H \cong G - uv + uw$. For example, if $F \cong K_{1,3}$, then the graph G of Figure 4 can be $K_{1,3}$-rotated into H and H′.

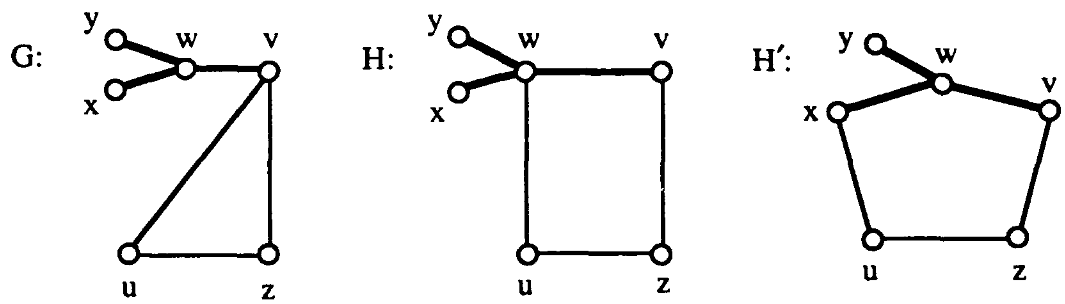


Figure 4

More generally, we say that a graph G can be *F-transformed* into H if either (1) $G \cong H$ or (2) there exists a sequence $G \cong G_0, G_1, \ldots, G_n \cong H$ of graphs such that, for $i = 0, 1, \ldots, n-1$, the graph G_i can be F-rotated into G_{i+1}. For instance, the graph G of Figure 5 *cannot* be $K_{1,4}$-rotated into H, but G *can* be $K_{1,4}$-transformed into H.

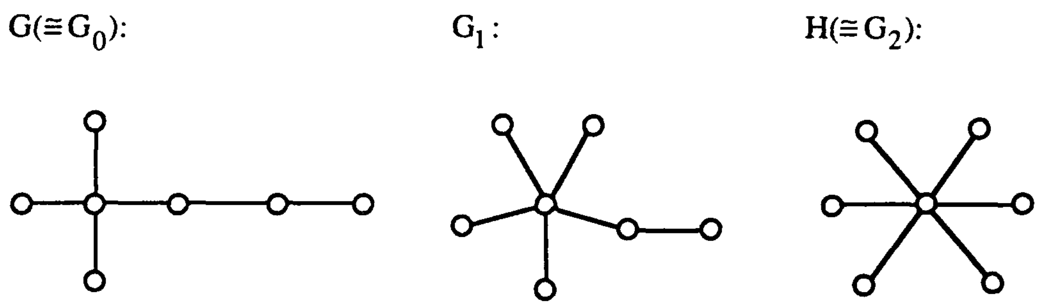


Figure 5

Observe that $\overline{K}_2$-rotation and K_2-rotation are edge rotation and edge slide, respectively. Clearly, if a graph G can be F-transformed into a graph H, then G and H have the same order, same size, and both contain a subgraph isomorphic to F. Unfortunately, the converse is not true, in general. For instance, the graphs G and H of Figure 6 have the same order and same size, and both G and H contain a subgraph isomorphic to C_4, but G cannot be C_4-transformed into H. In fact, G can be C_4-transformed only into itself.

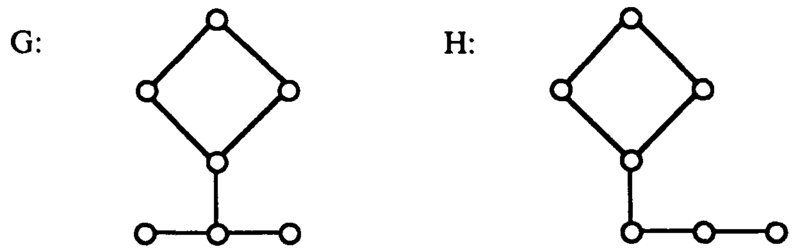


Figure 6

One may ask the question: What are necessary and sufficient conditions for one of two graphs G and H to be F-transformed into the other? We have already seen the answer to this question if $F \cong \overline{K}_2$ or $F \cong K_2$. The following results are due to Jarrett.

Proposition 9 Let F be any nontrivial connected graph. If a connected graph G can be F-transformed into a graph H, then H is connected.

Corollary 9a Let F be a nontrivial connected graph. A graph G can be F-transformed into a graph H if and only if the graph G has components $G_1, G_2, \ldots, G_k$, the graph H has components $H_1, H_2, \ldots, H_k$, and G_i can be F-transformed into H_i for every i $(1 \leq i \leq k)$.

Proposition 10 Let F be a connected graph of order p' with $\delta(F) = 1$, and let G and H be two (nonisomorphic) connected (p, q) graphs, each containing an induced subgraph isomorphic to F. Then G can be F-transformed into H.

Next we show that if $\delta(F) > 1$ or F is not an induced subgraph of G or H, the result does not necessarily hold. For example, for the graph F of Figure 7 we have $\delta(F) = 2 > 1$. Although the graphs G and H contain F as an induced subgraph, G cannot be F-transformed into H. In fact, G cannot be F-transformed into any graph different from G.

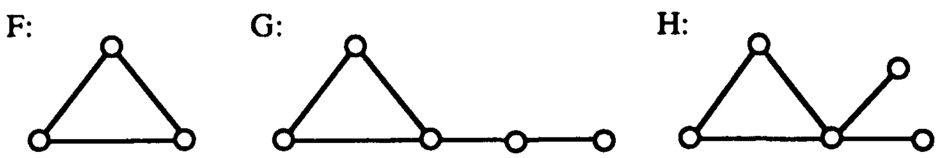


Figure 7

For the graph F of Figure 8 we have $\delta(F) = 1$, but F is *not* an induced subgraph of G, and H cannot be obtained from G by an F-transformation. Indeed, G can be F-transformed only into itself and graphs G' and G'' of Figure 8.

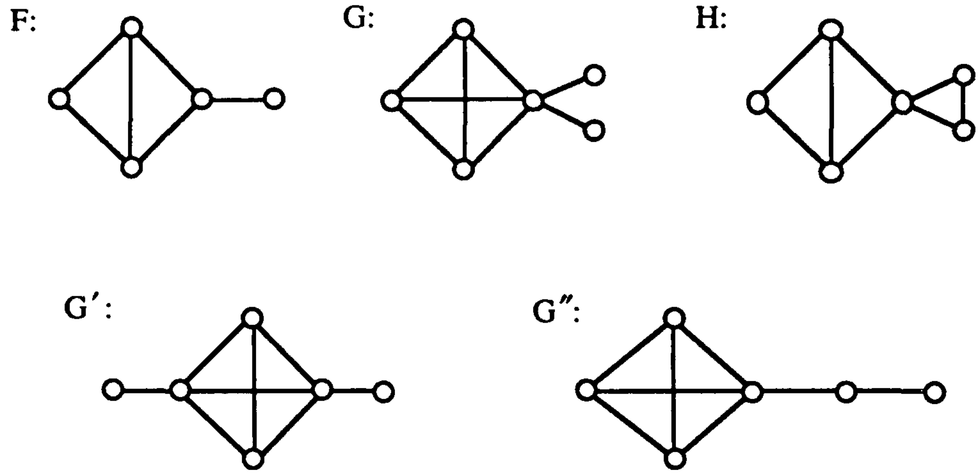


Figure 8

With each F-transformation another metric can be defined. Let F be a graph of order $p' \geq 2$ and let $\mathcal{S}$ be a set of (p, q) graphs such that for every pair G, H of graphs in $\mathcal{S}$, the graph G can be F-transformed into H. The *F-distance* F-d(G, H) between G and H is defined as the minimum number of F-rotations necessary to transform G into H. For the graphs F, G, and H of Figure 9, we have F-d(G, H) = 2.

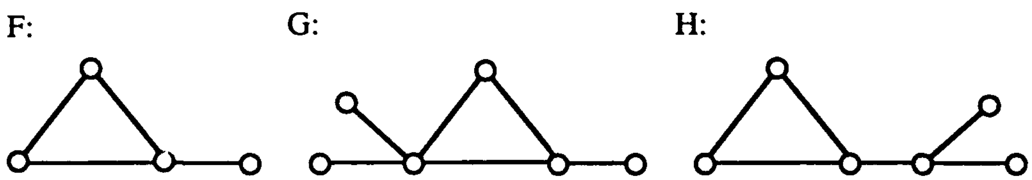


Figure 9

If a graph G cannot be F-transformed into a graph H we set F-d(G, H) = ∞. It is obvious that $\overline{K}_2$-d(G, H) $\leq$ F-d(G, H).

Proposition 11 Let F be a graph of order $p\ (\geq 2)$ and let n be a nonnegative integer. Then there exists a pair G_1, G_2 of graphs such that F-d(G_1, G_2) = n.

Let F be a graph of order $p'\ (\geq 2)$ and let $\mathcal{S}$ be a set of (p, q) graphs, each containing a subgraph isomorphic to F. Then the *F-distance graph* $\mathcal{D}_F(\mathcal{S})$ of $\mathcal{S}$ is that graph whose vertex set is $\mathcal{S}$ and in which two vertices G and H are adjacent if and only if F-d(G, H) = 1. For example, if $F \cong K_3$ and $\mathcal{S}$ is the set of graphs G_1, G_2, G_3, and G_4 shown in Figure 10, then $\mathcal{D}_F(\mathcal{S}) \cong K_4 - e$.

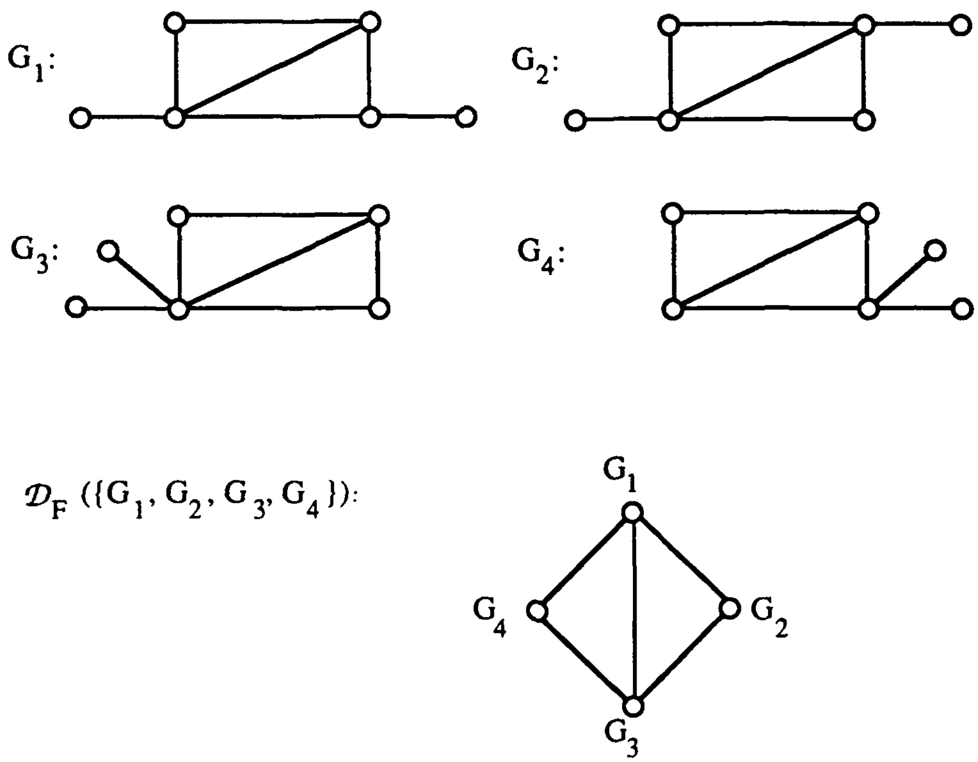


Figure 10

Proposition 12 Let F be a nontrivial connected graph. Then every graph is an F-distance graph.

Subgraph Distance

Let G_1 and G_2 be edge-induced subgraphs of the same size in a graph G. The subgraph G_2 can be obtained from G_1 by an *edge rotation* if there exist distinct vertices u, v, and w such that $uv \in E(G_1)$, $uw \notin E(G_1)$, and $G_2 = G_1 - uv + uw$. More generally, G_1 can be *r-transformed* into G_2 if $G_1 = G_2$ or G_2 can be obtained from G_1 by a sequence of edge rotations. It was shown by Chartrand, Johns, Novotny, and Oellermann that every edge-induced subgraph of a connected graph G can be r-transformed into any edge-induced subgraph of G having the same size. The *edge rotation distance* $d_r(G_1, G_2)$ between G_1 and G_2 is the minimum number of edge rotations required to r-transform G_1 into G_2. For the graph G of Figure 11, the subgraph G_3 can be obtained from G_1 by an edge rotation so that $d_r(G_1,G_3) = 1$. On the other hand, G_3 cannot be obtained from G_2 by an edge rotation, but G_3 can be obtained from G_2 by an r-transformation and $d_r(G_2, G_3) = 3$.

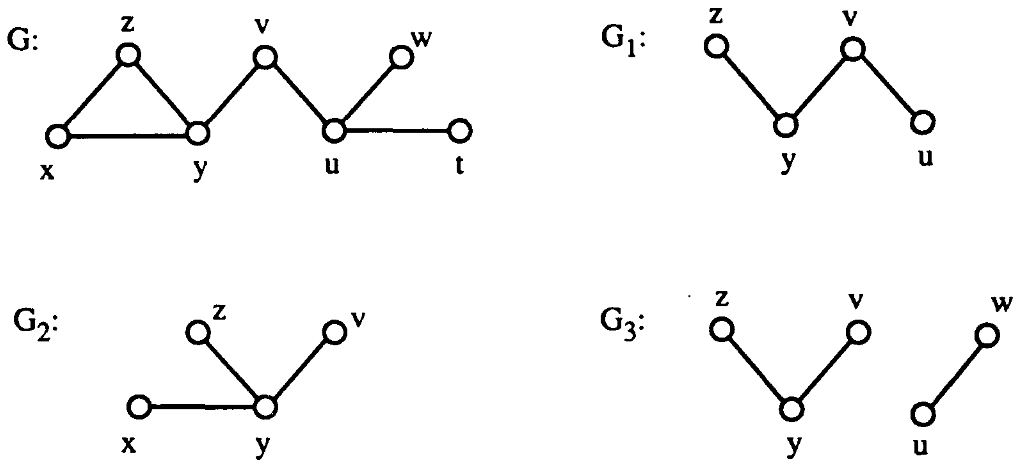


Figure 11

One can also define a subgraph transformation based on edge slide. Let G_1 and G_2 be two edge-induced subgraphs of the same size in G. We say that G_2 can be obtained from G_1 by an *edge slide* if there exist distinct vertices u, v, and w of G such that $uv \in E(G_1)$, $uw \notin E(G_1)$, $vw \in E(G)$, and $G_2 = G_1 - uv + uw$. For example, for the graph G of Figure 12, the subgraph G_2 can be obtained from G_1 by an edge slide. More generally, we say that G_1 can be *s-transformed* into G_2 if either (1) $G_1 = G_2$ or (2) G_2 can be obtained from G_1 by a sequence of edge slides.

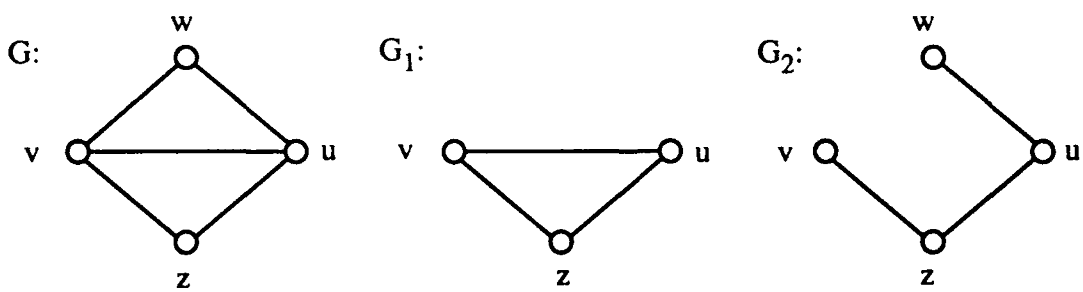


Figure 12

As we mentioned earlier, for every pair H, H′ of edge-induced subgraphs of the same size in a connected graph G, the subgraph H can be r-transformed into H′. Unfortunately, this is not the case for s-transformations. For example, if G is the graph of Figure 13, then H cannot be s-transformed into H′. In fact, H can be s-transformed only into itself.

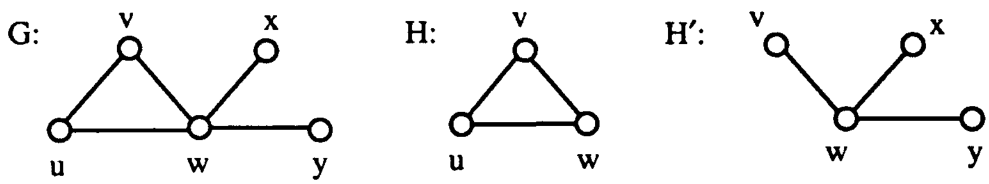


Figure 13

Let e and f be edges of a graph G. A *triangular e-f walk* of G is a finite, alternating sequence $e = e_0, T_1, e_1, T_2, \ldots, e_{n-1}, T_n, e_n = f$ of edges and triangles such that e_{i-1} and e_i belong to T_i $(1 \le i \le n)$. A *triangular* e-f *path* is a triangular e-f walk in which no edges or triangles are repeated. The number n of triangles in the triangular path is called its *length*. In the graph G of Figure 14 there exists a triangular e-f path (with $T_i = \langle\{e_{i-1}, e_i\}\rangle$, i = 1, 2, 3), but there is no triangular e-g path.

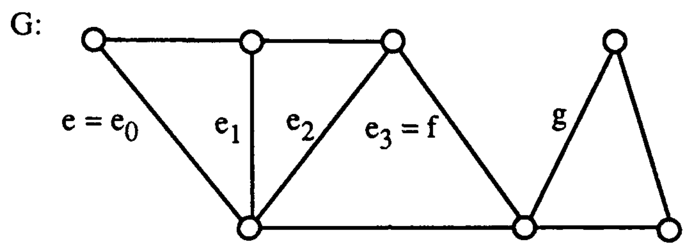


Figure 14

It is straightforward to show that every triangular e-f walk in a graph contains a triangular e-f path.

Observe that for every two edges e and e′ of a triangle T, the subgraph $\langle\{e'\}\rangle$ can be obtained from $\langle\{e\}\rangle$ by an edge slide. Therefore, if in a graph G there exists a triangular e-f path, then $G_1 = \langle\{e\}\rangle$ can be s-transformed into $G_2 = \langle\{f\}\rangle$.

Whenever edges e and f belong to a 3-cycle in G, we denote this triangle by T(e, f) and call it a *slide induced triangle*. With every edge slide there is associated a unique triangle T, namely, if $G_2 = G_1 - e + f$, then T = T(e, f). These observations are useful in proving the following result by Jarrett.

Proposition 13 Let G be a connected graph of size $q \ge 1$, and let q' be an integer with $1 \le q' \le q$. For every pair G_1, G_2 of subgraphs of G having size q', the subgraph G_1 can be s-transformed into G_2 if and only if every two edges of G are connected by a triangular path.

Triangular Line Graphs

For a given graph G, we define its *triangular line graph* $\mathcal{T}(G)$ as that graph with vertex set $E(G)$ such that two vertices e and f of $\mathcal{T}(G)$ are adjacent if and only if $T(e, f)$ is a triangle of G. For $G \cong K_4 - e$, the graph $\mathcal{T}(G)$ is shown in Figure 15.

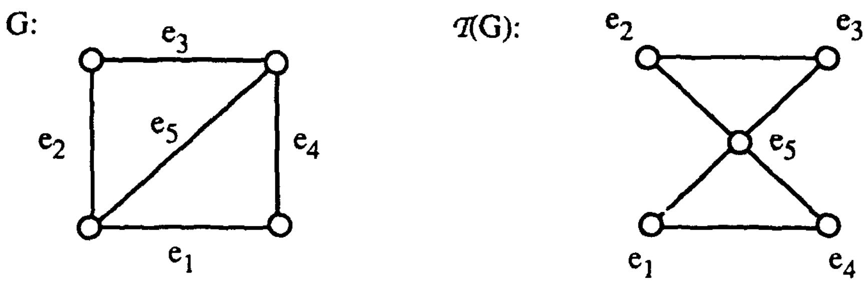


Figure 15

It follows from the definition that $\mathcal{T}(G)$ is a spanning subgraph of the line graph $L(G)$. The next result is perhaps less obvious. These are due to Jarrett.

Proposition 14 Let G be a connected graph of order $p \geq 2$. Then $\mathcal{T}(G) = L(G)$ if and only if $G \cong K_p$.

Proposition 15 Let G be a connected nontrivial graph. For every pair G_1, G_2 of edge-induced subgraphs of G having size 1, the subgraph G_1 can be s-transformed into G_2 if and only if $\mathcal{T}(G)$ is connected.

Corollary 15a Let G be a nontrivial connected graph. Then for every pair G_1, G_2 of edge-induced subgraphs of G having the same size, G_1 can be s-transformed into G_2 if and only if $\mathcal{T}(G)$ is connected.

For integers $n \geq 2$, the *nth iterated triangular line graph* $\mathcal{T}^n(G)$ of a graph G is defined to be $\mathcal{T}(\mathcal{T}^{n-1}(G))$, where $\mathcal{T}^1(G)$ denotes $\mathcal{T}(G)$ and $\mathcal{T}^{n-1}(G)$ is assumed to be nonempty. Clearly, $\mathcal{T}^n(G)$ is a subgraph of the nth iterated line graph $L^n(G)$ of G. In fact, for $n = 1$, $\mathcal{T}^1(G) = \mathcal{T}(G)$ is a spanning subgraph of $L^1(G) = L(G)$.

Note that every triangle T in G gives rise to a triangle T' in $\mathcal{T}(G)$ with a one-to-one correspondence between the edges of T and the vertices of T'. Moreover, if T_1 and T_2 are two triangles of G, then the corresponding triangles T_1' and T_2' of $\mathcal{T}(G)$ are edge-disjoint. For suppose, to the contrary, that T_1' and T_2' have an edge in common or, equivalently, T_1' and T_2'

have two common vertices, say e and f. Necessarily, e and f are common edges of T_1 and T_2, which implies that $T_1 = T_2$. Thus $\mathcal{T}(G)$ has at least as many triangles as G has. We show that only for K_4-free graphs G are the number of triangles in G and $\mathcal{T}(G)$ equal.

Proposition 16 Let t(G) denote the number of triangles in a graph G. Then $t(G) = t(\mathcal{T}(G))$ if and only if G is K_4-free.

Proposition 17 Let G be a K_4-free graph. Then $\mathcal{T}^n(G) \cong \mathcal{T}^2(G)$, for $n \geq 2$.

The previous result does not hold for a graph $G \cong K_4$. However, $\mathcal{T}^3(K_4) \cong 8K_3$ and, therefore, for $n \geq 3$ we have $\mathcal{T}^n(K_4) \cong \mathcal{T}^3(K_3)$. The graphs $G \cong K_4$ and $\mathcal{T}^i(K_4)$, $1 \leq i \leq 3$, are shown in Figure 16.

Conjecture For every graph G containing at least one triangle, there exists an integer $k > 0$, such that for $n \geq k$, $T^n(G) \cong T^k(G)$.

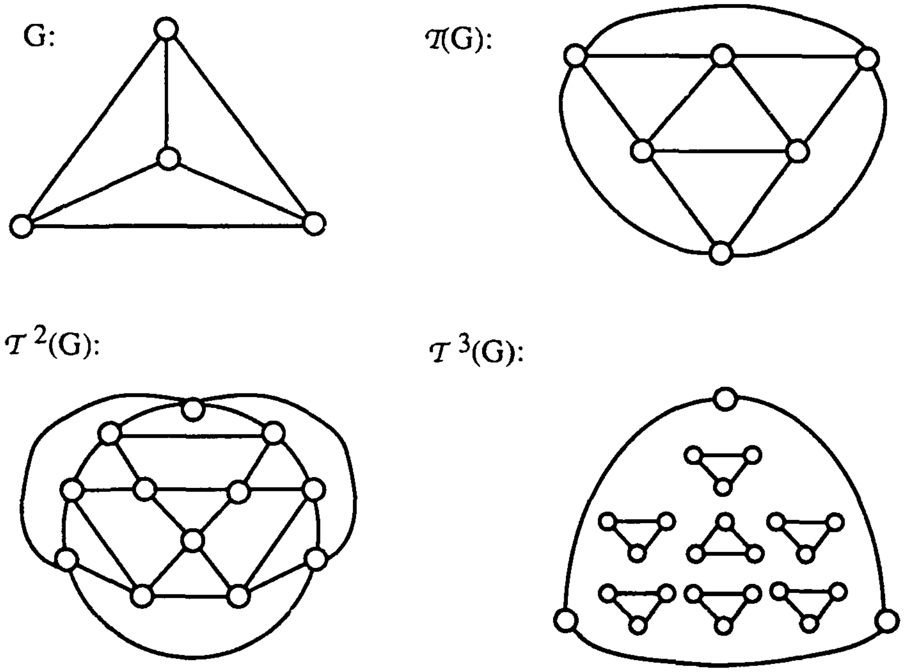


Figure 16

n–Subgraph Distance Graphs

The n-subgraph distance graphs were introduced by Chartrand, Hevia, Jarrett, Saba, and VanderJagt. Let G be a graph of size q (≥ 1) and let n be an integer with $1 \leq n \leq q$. The *n-subgraph distance graph* $L_n(G)$ of G is that graph whose vertices correspond to the edge-

induced subgraphs of size n in G and where two vertices of $L_n(G)$ are adjacent if and only if the edge rotation distance between corresponding subgraphs is 1. It is convenient to label the vertices of $L_n(G)$ by the edge sets of the corresponding subgraphs or simply by listing the edges. Each edge in a vertex label is called a *coordinate*. Since the coordinates are elements of a set, the order in which the coordinates of a vertex are listed is irrelevant. For example, if a vertex of $L_n(G)$ corresponds to the subgraph of G induced by the edge set $\{e_1, e_2, ..., e_r\}$, then we may label this vertex as $e_1, e_2, ..., e_n$ or $e_i \cup X$, where $X = \{e_j \mid 1 \le j \le n, j \ne i\}$, or simply as e_iX. For the graph $G \cong K_1 + (K_1 \cup K_2)$ of Figure 17, the graphs $L_i(G)$, $i = 1, 2, 3, 4$, are shown.

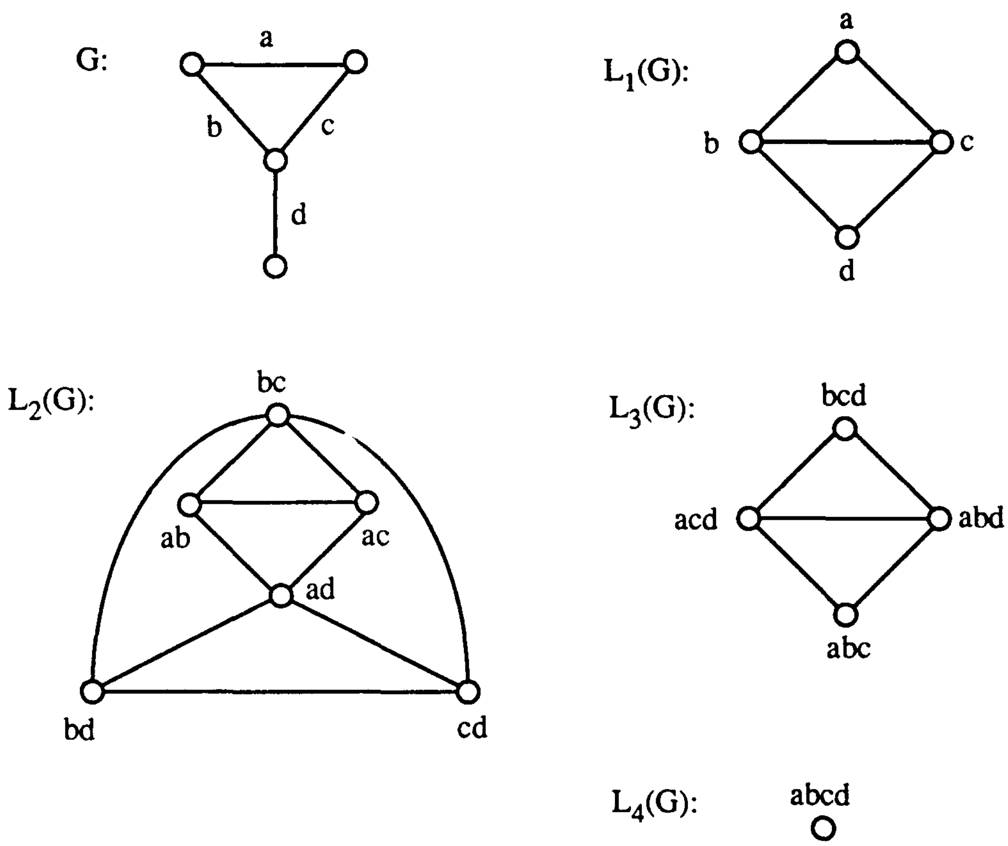


Figure 17

Observe that $L_4(G) \cong K_1$ for the graph G of Figure 17. In general, $L_q(G) \cong K_1$ for a graph G of size q. Furthermore, $L_1(G) \cong L_3(G)$ for the graph G of Figure 17. This fact illustrates the following result.

Proposition 18 Let G be a graph of size q, and let n be an integer with $1 \le n < q$. Then $L_n(G) \cong L_{q-n}(G)$.

The graphs $L_n(G)$, $1 \le n \le q = E(G)$, are also called generalized line graphs since the 1-subgraph distance graph $L_1(G)$ is the line graph of G. We shall also refer to these graphs as *n-subgraph rotation distance graphs* to distinguish them from the n-subgraph slide distance graphs, which we are about to describe. We begin with the definition of n-subgraph slide distance.

Let G be a graph of size q (≥ 1), and let G_1 and G_2 be two edge-induced subgraphs of G having the same size n ($1 \le n \le q$). We define the *n-subgraph slide distance* $d_s(G_1, G_2)$ between G_1 and G_2 as the smallest nonnegative integer k for which there exists a sequence $H_0, H_1, \ldots, H_k$ of subgraphs of G such that $G_1 \cong H_0$, $G_2 \cong H_k$ and, for $i = 1, 2, \ldots, k$, H_i can be obtained from H_{i-1} by an edge slide. If no such k exists, we define $d_s(G_1, G_2) = \infty$. If $G \cong K_4 - e$, and G_1 and G_2 are two subgraphs of G shown in Figure 18, then $d_s(G_1, G_2) = 2$.

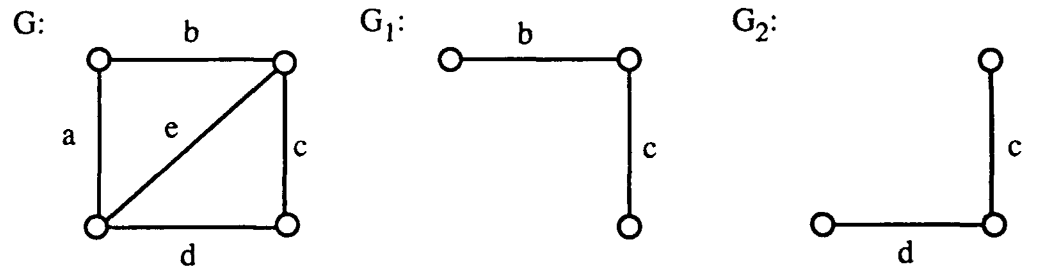


Figure 18

We define the *n-subgraph slide distance graph* $S_n(G)$ of G as the graph whose vertices correspond to the edge-induced subgraphs of size n and where two vertices G_1 and G_2 of $S_n(G)$ are adjacent if and only if $d_s(G_1, G_2) = 1$. It is straightforward to see that $S_1(G) = \mathcal{T}(G)$, and, therefore, $S_1(G)$ is a spanning subgraph of $L_1(G)$. In general, for $n \ge 1$, $S_n(G)$ is a spanning subgraph of $L_n(G)$. For the graph $G \cong K_4 - e$, the graphs $S_i(G)$, $1 \le i \le 5$, are shown in Figure 19.

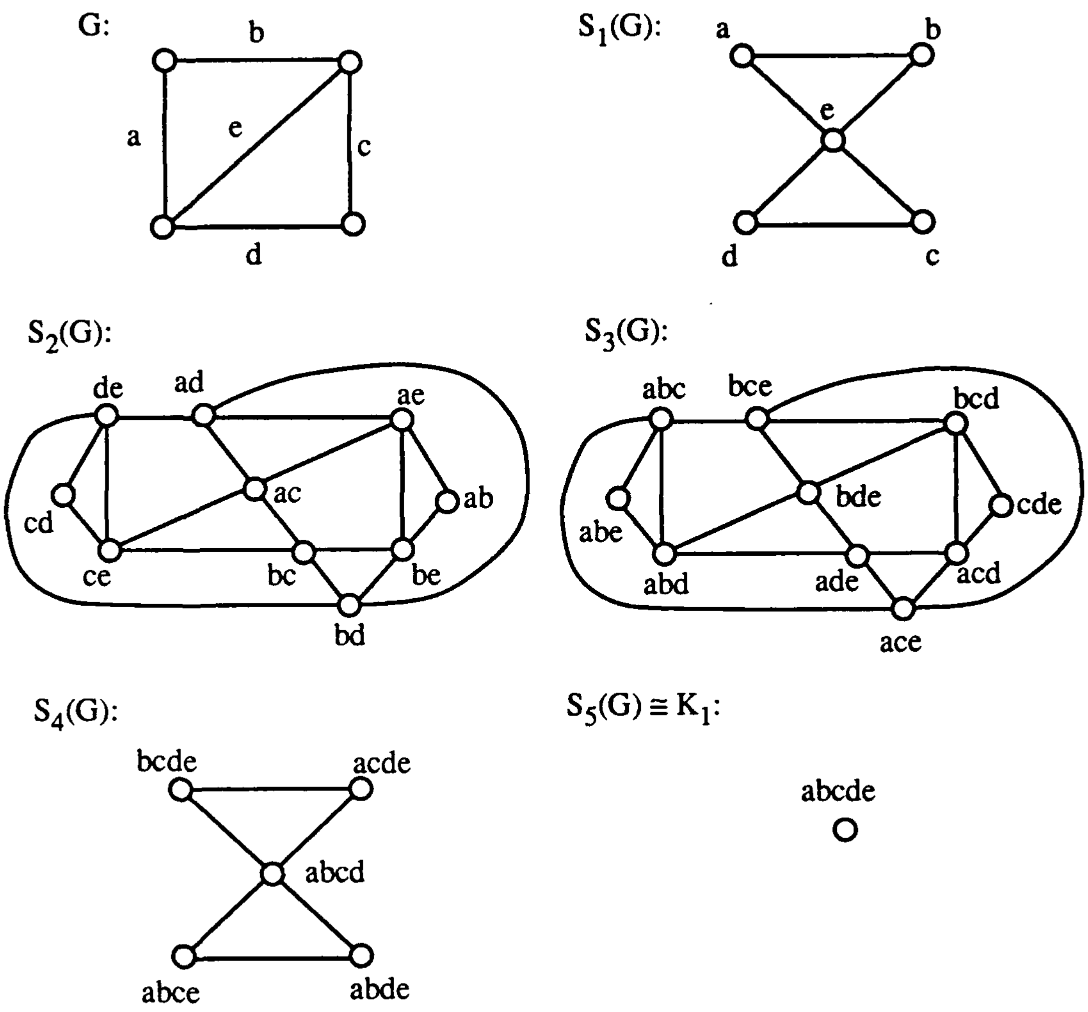


Figure 19

Observe that for the graph $G \cong K_4 - e$, we have $S_2(G) \cong S_3(G)$ and $S_1(G) \cong S_4(G)$. This fact can be generalized as follows.

Proposition 19 Let G be a graph of size q (≥ 1) and let n be an integer with $1 \leq n < q$. Then $S_n(G) \cong S_{q-n}(G)$.

Subgraph Distance for Subgraphs of the Same Order

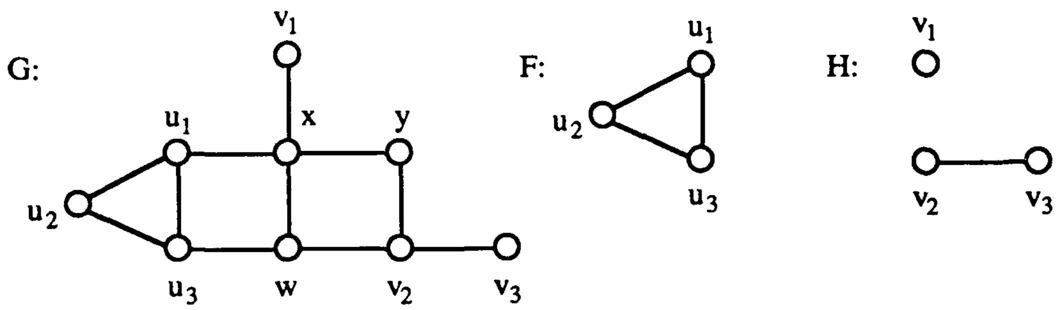
For a connected graph G of order p and an integer n such that $1 \leq n \leq p$, let F and H be induced subgraphs of G of order n. We define a pairing π from the set V(F), say $\{v_1, v_2, \dots, v_n\}$, to the set V(H) as a one-to-one correspondence that associates a vertex of V(F) with one of V(H). The distance induced by π between F and H is defined as

$$d_\pi(F, H) = \sum_{i=1}^{n} d(v_i, \pi(v_i))$$

and the subgraph distance between F and H is

$$d(F, H) = \min_{\pi} d_\pi(F, H).$$

This concept was introduced by Chartrand, Johns, Novotny, and Oellermann. Observe that if F consists of a single vertex, say u, and H consists of a single vertex, say v, then d(F, H) = d(u, v). Thus d(F, H) is a generalized distance defined in terms of subgraphs. As an example, Figure 20 gives a connected graph G, two induced subgraphs F and H of G, a listing of all pairings from V(F) to V(H), and d(F, H).



Pairings	u_i	v_j	$d(u_i, v_j)$	$d\pi_k(F, H)$
π_1	u_1 u_2 u_3	v_1 v_2 v_3	2 3 3	8
π_2	u_1 u_2 u_3	v_1 v_3 v_2	2 4 2	8
π_3	u_1 u_2 u_3	v_2 v_1 v_3	3 3 3	9
π_4	u_1 u_2 u_3	v_2 v_3 v_1	3 4 3	10
π_5	u_1 u_2 u_3	v_3 v_1 v_2	4 3 2	9
π_6	u_1 u_2 u_3	v_3 v_2 v_1	4 3 3	10

$d(F, H) = 8$

Figure 20

n-Vertex Graphs

Let G be a connected graph of order p and let F and H be subgraphs of order n with $1 \leq n \leq p-1$. Then $d(F, H) = 1$ if and only if there exist adjacent vertices $u \in V(F)$ and $v \in V(H)$ such that $V(F-u) = V(H-v)$. We define the *n-vertex graph* of G as that graph G_n whose vertices are the induced subgraphs of order n in G and two vertices F and H of G_n are adjacent if and only if $d(F, H) = 1$. A graph and its 2-vertex graph are shown in Figure 21.

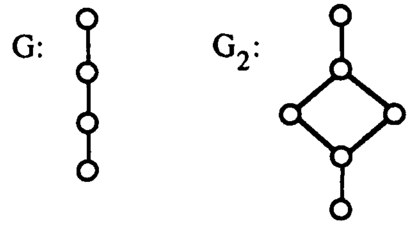


Figure 21

The optimality of clustering and monotone optimal assemblies
by
Uriel G. Rothblum

The purpose of the study is to present recent results which provide sufficient conditions for the optimality of clustered partitions. In particular, we introduce a formal definition of clustering by calling a partition clustered if convex hulls of the sets of the partition are disjoint. It is shown that if given vectors $A1, \ldots, An$ in Rk are to b partitioned into m groups, of predetermined sizes, so as to maximize an objective which is a quasi-convex function of the sums of the vectors in each set, then a clustered optimal partition exists. Further, if quasi-convexity is replaced by strict quasi-convexity, every optimal partition is clustered. Computational implications of the results are discussed.

The techniques we developed enabled us to determine sufficient conditions for the optimality of monotone assemblies. Here we consider the problem of identifying multipartitions which occur when items of different types are to be partitioned into sets, for example, in system assembly, components of different type are assigned to modules which compose the system. We show that if the system is coherent and if the components in each module are in series, there is a reliability maximizing assembly which is monotone, i.e., it has a single module which gets the best parts of each type (according to its specification), there is another module which gets the second best parts of each type (according to its specification), and so on, till finally, there is one module which gets the worst parts of each type (according to its specification).

THE OPTIMALITY OF CLUSTERING

OR

OPTIMAL PARTITIONS HAVING DISJOINT CONVEX OR CONIC HULLS

U. ROTHBLUM

RUTCOR

RUTGERS UNIVERSITY

v

OUTLINE

- PARTITIONING PROBLEMS
- MAIN RESULT ABOUT OPTIMAL PARTITIONS
- COMPUTATIONAL SIGNIFICANCE
- EXAMPLES - NEYMAN PEARSON LEMMA
  LOCATION PROBLEMS
- ONGOING RESEARCH

CLASSIFICATION AND CLUSTERING

MANY PROBLEMS HAVE THE FOLLOWING STRUCTURE:

- PARTITION A NUMBER OF ENTITIES INTO GROUPS
- EACH ENTITY HAS SEVERAL NUMERICAL CHARACTERISTICS
- AN OBJECTIVE IS ASSOCIATED WITH EACH PARTITION

EXAMPLES: STATISTICS - HYPOTHESIS TESTING

INENTORY GROUPING

ASSEMBLY OF PARTS IN A SYSTEM

LOCATION PROBLEMS

CLASSIFICATION PROBLEMS

COMMON PRACTICE - CLUSERING

GROUP TOGETHER ENTITIES WITH SIMILAR CHARACTERISTICS

PARTITIONING PROBLEMS - DEFINITION

$a^1, \dots, a^n$ (DISTINCT) VECTORS IN R^k

$S_1, \dots, S_m$ PARTITION OF $\{1, \dots, n\}$

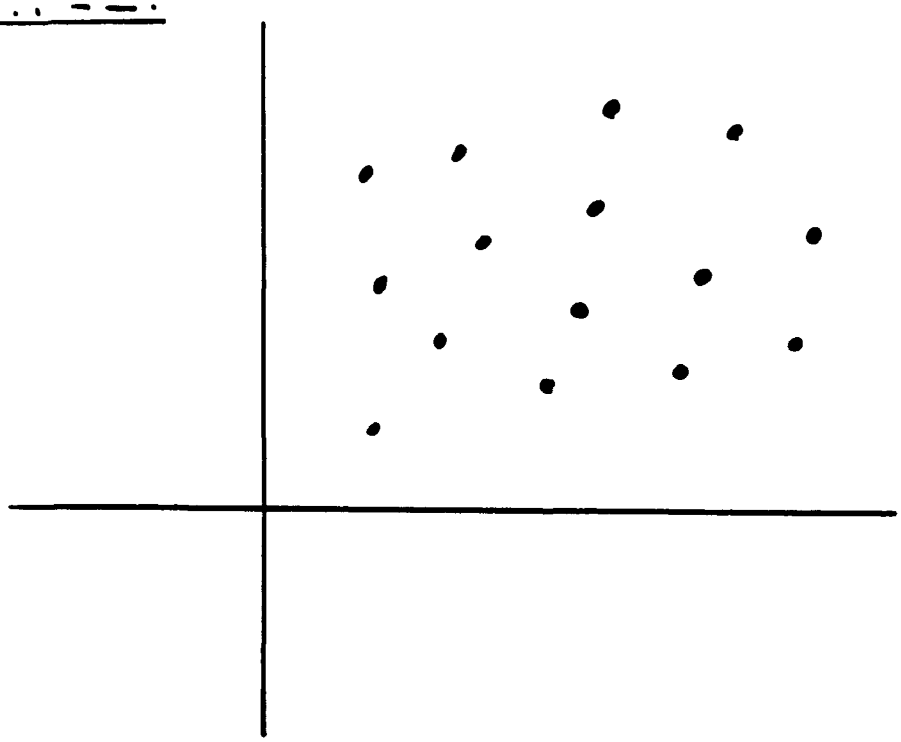
$h(S_1, \dots, S_m)$ OBJECTIVE

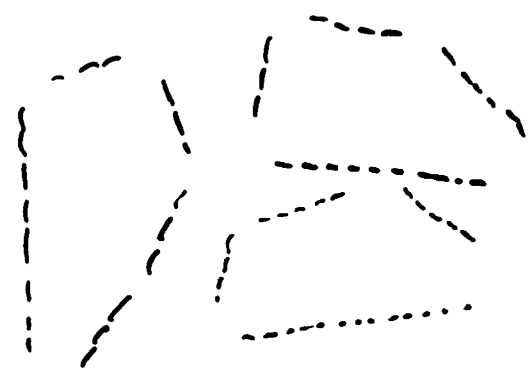
GOAL IS TO MINIMIZE $h(S_1, \dots, S_m)$

RESTRICTED PROBLEM: $|S_j|$ GIVEN APRIORI

UNRESTRICTED PROBLEM: $|S_j|$ UNRESTRICTED

EXAMPLE:





CONVEX HULLS ARE DISJOINT

MAIN THEOREM

$h(S_1, \dots, S_m)$ IS CALLED QUASI CONCAVE IF

$$h(S_1, \dots, S_m) = f\Big(\sum_{i \in S_1} a^i, \dots, \sum_{i \in S_m} a^i\Big)$$

AND f IS QUASI-CONCAVE IN mk VARIABLE.

EXAMPLE: HYPOTHESIS TESTING

$\{1, \dots, n\}$ OUTCOME OF EXPERIMENT

$p_i = Pr(i \mid H_0)$

$q_i = Pr(i \mid H_1)$

PARTITION: PARTITION $\{1, \dots, n\}$ INTO ACCEPT H_0 AND REJECT H_0 GROUPS

OBJECTIVE: PROBABILITY OF FIRST TYPE ERROR: $\sum_{i \in A} q_i Pr(H$

PROBABILITY OF SECOND TYPE ERROR: $\sum_{i \in R} p_i Pr(H_0)$

$$h(A,R) = \min \; c_1 \Big(\sum_{i \in A} q_i\Big) Pr(H_1) + c_2 \Big(\sum_{i \in R} p_i\Big) Pr(H_0)$$

HERE f IS LINEAR

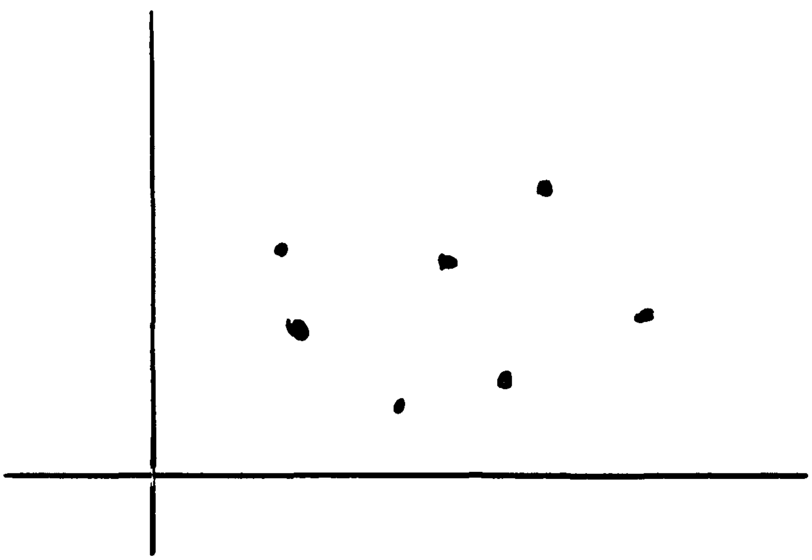
THEOREM 1:

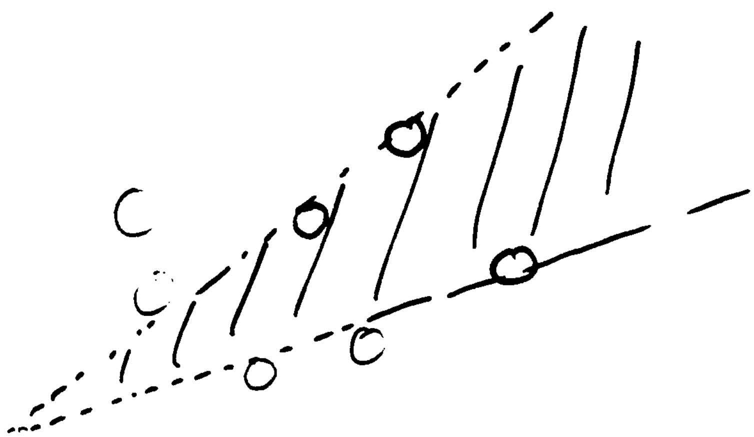
IF h IS QUASI-CONCAVE THEN THE RESTRICTED PROBLEM HAS AN OPTIMAL PARTITION WITH DISJOINT CONVEX HULLS

THEOREM 2:

IF h IS QUASI-CONCAVE THEN THE UNRESTRICTED PROBLEM HAS AN OPTIMAL PARTITION WITH DISJOINT CONIC HULLS

DISJOINT CONIC HULLS:





6d

7

IDEA OF PROOF:

- CONSIDER THE POLYTOPE IN R^{mk}

$$C = \mathrm{CONV}\left\{\left(\sum_{i\in S_1} a^i, \dots, \sum_{i \in S_m} a^i\right) : S_1, \dots, S_m \text{ PARTITIC } 1, \dots, n\right.$$

IT IS PROVED THAT THE EXTREME POINTS OF THIS POLYTOPE CORRESPOND TO PARTITIONS WITH THE DESIRED PROPERTY

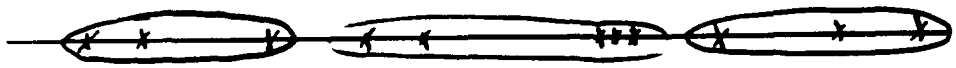
NOTE: DESIRED PROPERTIES DO NOT CHARACTERIZE EXTREME POINTS

THEOREM 3.

IF f IS STRICTLY QUASI CONCAVE EVERY OPTIMAL PARTITION HAS THE DESIRED PROPERTY

COMPUTATIONAL SIGNIFICANCE

$k=1$: DISJOINT CONVEX HULLS

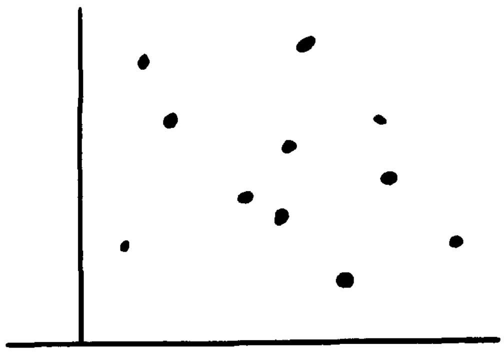


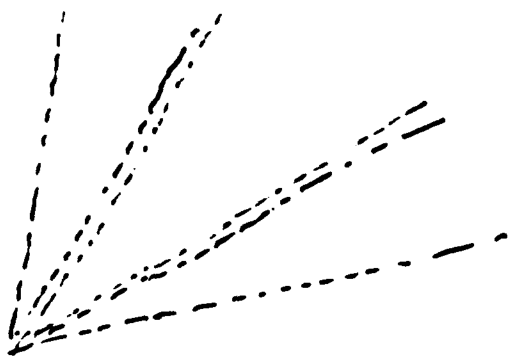
THESE ARE CALLED CONSECUTIVE PARTITIONS

\# OF CONSECUTIVE PARTITIONS: $O(n^{m-1})$

\# OF ALL PARTITIONS: m^n

$k=2$ - DISJOINT CONIC HULLS





ORDER VECTORS SO THAT

$$\frac{y_1}{x_1} \leq \frac{y_2}{x_2} \leq \dots \leq \frac{y_n}{x_n}$$

THEN THERE WILL BE AN OPTIMAL PARTITION

WHICH IS CONSECUTIVE

OF CONSECUTIVE PARTITIONS $O(n^{m-1})$

OF ALL PARTITIONS m^n

k. GENERAL, m=2, DISJOINT CONVEX HULLS

OF ALL PARTITIONS 2^n

OF STRUCTURED PARTITIONS $O\binom{n}{k}$

TYPICAL IN GENERAL

OF ALL PARTITIONS EXPONENTIAL IN n

OF STRUCTURED PARTITIONS POLYNOMIAL IN

OPEN PROBLEM:

FIND EFFICIENT METHODS FOR SYSTEMATIC ENUMERATION

SPECIAL CASES HAVE BEEN DONE

HYPOTHESIS TESTING - REVISITED

$$\min \; c_1\left(\sum_{i\in A} q_i\right) \Pr(H_1) + c_2\left(\sum_{i\in R} p_i\right) \Pr(H_0)$$

OUR THEOREM SHOWS THE EXISTENCE OF AN OPTIMAL PARTITION (A, R) WHICH IS CONSECUTIVE IN $\frac{p_i}{q_i}$

$\frac{p_i}{q_i} \geq \alpha$ ACCEPT

$\frac{p_i}{q_i} < \alpha$ REJECT

NEYMAN PEARSON LEMMA

LOCATION PROBLEMS - CIRCLE SEPARATIO

SUPPOSE $a^1,\ldots,a^n$ IN R^k ARE TO BE PARTITIONED INTO $S_1,\ldots,S_m$ SO AS TO MINIMIZE

$$\sum_m \sum_{i,j\in S_m} \|a^i - a^j\|^2$$

OBSERVATION:

$$\sum_m \sum_{i,j\in S_m} \|a^i - a^j\|^2 = 2\left[\sum_m |S_m|\left(\sum_{i\in S_m} \|a^i\|^2\right) - \left\|\sum_{i\in S_m} a^i\right\|^2\right]$$

$$= 2\sum_m \left(x_m y_m - \|z_m\|^2\right)$$

WHERE

$$x_m = \sum_{i\in S_m} \|a^i\|^2, \quad y_m = |S_m|, \quad z_m = \sum_{i\in S_m} a^i$$

ANALYSIS

IDENTIFY a^i WITH $\begin{pmatrix} \|a^i\|^2 \\ 1 \\ a^i \end{pmatrix}$

AND APPLY THEOREM 1

CONCLUSION: $\quad i \to \begin{pmatrix} \|a^i\|^2 \\ a^i \\ 1 \end{pmatrix}$

THERE IS AN OPTIMAL PARTITION SUCH THAT EACH PAIR OF SETS S_1 AND S_2 HAS SOME

$$\alpha \in R, \quad \beta \in R^k, \quad \gamma \in R$$

SUCH THAT

$$\alpha \|a^i\|^2 + \beta^T a^i + \gamma \geq \alpha \|a^j\|^2 + \beta^T a^j + \gamma$$

$$\forall i \in S_1, \; j \in S_2$$

OBSERVE:

IF $\alpha > 0$: $\quad \|a^i + \frac{\beta}{2\alpha}\|^2 \geq \|a^j + \frac{\beta}{2\alpha}\|^2$

CIRCLE SEPARATION

IF $\alpha = 0$: HYPERPLANE SEPARATION

IF $\alpha < 0$: CIRCLE SEPARATION WITH S_1 AND S_2 REVERSED

OPTIMALITY OF MONOTONE ASSEMBLIES FOR COHERENT SYSTEMS HAVING SERIES MODULES

JOINT WORK WITH

FRANK K. HWANG

THE MODEL

PARTS OPERATIVE/INOPERATIVE

MODULES SERIES STRUCTURE

SYSTEM ARBITRARY COHERENT STRUCTU

1) STATE OF SYSTEM DEPENDS ON STATES OF MODULES

$$J: \{0,1\}^m \longrightarrow \{0,1\}$$

2) MONOTONE

3) NO REDUNDENT MODULE

DATA

m # OF MODULES

n_{ui} # OF PARTS OF TYPE u NEEDED FOR MODULE i

$n_u = \sum_i n_{ui}$ # OF PARTS OF TYPE u AVAILABLE

t # OF PART TYPES

r_{uk} RELIABILITY OF PART k OF TYPE u

W.L.O.G.

$$r_{u1} \geq r_{u2} \geq \cdots \geq r_{un_u}$$

ASSEMBLIES

AN ASSEMBLY π CONSISTS OF SETS

$$\{\pi_{ui} : u = 1, \ldots, t, \; i = 1, \ldots, m\}$$

SUCH THAT: (a) $|\pi_{ui}| = n_{ui}$

(b) $\pi_{u1}, \ldots, \pi_{um}$ PARTITIONS $\{1, \ldots, n_u\}$

MODULE AND SYSTEM RELIABILITY

SUPPOSE ASSEMBLY $\pi = \{\pi_{ui}\}$ IS USED

RELIABILITY OF MODULE i:

$$r(\pi)_i = \prod_{u=1}^{t} \left[\prod_{k \in \pi_{ui}} \rho_{uk} \right]$$

SYSTEM OPERABILITY FUNCTION

$$J : \{0,1\}^m \to \{0,1\}$$

SYSTEM RELIABILITY IF r_i IS THE RELIABILITY OF MODULE i, $1 \le i \le m$

$$f : [0,1]^m \to [0,1]$$

$$f(r) = \sum_{s \in \{0,1\}^m} J(s) \left[\prod_{s_i = 0} (1 - r_i) \right] \left[\prod_{s_i = 1} r_i \right]$$

SYSTEM RELIABILITY UNDER π

$$R(\pi) = f[r(\pi)]$$

OPTIMAL ASSEMBLY

OPTIMALITY OF MONOTONE ASSEMBLIES

MONOTONE ASSEMBLY: AN ASSEMBLY π IS CALLED MONOTONE IF THE MODULES CAN BE RENUMBERED SO THAT

$$i < j,\ k \in \pi_{ui},\ p \in \pi_{uj} \Rightarrow k < p$$

FOR ALL u.

THEOREM: THERE EXISTS AN OPTIMAL ASSEMBLY WHICH IS MONOTONE

THEOREM: IF

ⓐ $\Phi \geq r_{u1} \geq r_{u2} \geq \dots \geq r_{un_u} \geq 0$

ⓑ NO TWO MODULES ARE IN SERIES

THEN EVERY OPTIMAL ASSEMBLY IS MONOTONE.

An Ordinal Approach to Consensus Functions

G. D. Crown and M. F. Janowitz

Abstract. Let G be a finite nonempty set. A consensus function can be viewed as a mapping $F:G^k \longmapsto G$. For $\pi \in G^k$, $F(\pi)$ is often constructed from certain building blocks. For a family of relations on a set X, these might be the pairs (x,y) that constitute the output relation. In its most general setting, one can work with a triple (G,S,γ) where S is the set of building blocks and $\gamma:G \longmapsto \mathcal{P}(S)$ is a one-one function. A *neutral* consensus function F is characterized by the existence of a family $\mathcal{D}_F$ of subsets of $V_k = \{1,2,\ldots,k\}$ having the property that for any k-tuple $\pi \in G^k$, $s \in \gamma(F(\pi))$ if and only if $\{i: s \in \gamma(g_i)\} \in \mathcal{D}_F$. Assuming that $V_k \in \mathcal{D}_F$, and $\varnothing \notin \mathcal{D}_F$, we investigate conditions on γ that are equivalent to: $A \in \mathcal{D}_F$ implies $V_k \backslash A \notin \mathcal{D}_F$; $A \notin \mathcal{D}_F$ implies $V_k \backslash A \in \mathcal{D}_F$; $\mathcal{D}_F$ is an order filter (principal filter, ultrafilter) of V_k. The work is in progress, but here are some sample results: Theorem. Assume G has elements g,h such that $\gamma(g) \cup \gamma(h)$ is not of the form $\gamma(v)$, while $\gamma(g) \cap \gamma(h) = \gamma(w)$ for some $w \in G$. Than for any neutral F, it is true that $A,B \in D_F$ implies $A \cap B \neq \varnothing$. (Included here would be any nondistributive lattice, as well as any meet semilattice that is not a lattice). There are also some general results that imply the following: If G is a lattice having a pair of comparable sup-irreducibles, then $\mathcal{D}_F$ is an order filter of V_k. If G is an atomistic lattice having distinct atoms a,b,c such that $c \leq a \vee b$, then $\mathcal{D}_F$ is a principal filter (so any neutral consensus method is an oligarchy). If G is an atomistic join semilattice that is not a lattice having distinct atoms a,b,c such that $c \leq a \vee b < 1$, or if G is a nondistributive join semilattice that is not a lattice having at least 4 atoms, such that every line contains exactly 2 points, then any neutral consensus method is a dictatorship.

ORDINAL CONSENSUS METHODS

G. D. Crown and M. F. Janowitz

(Preliminary Report)

What is a consensus method?

G = finite set of objects
(the objects to be summarized)

V = indexing set (the "voters")
(fixed, possibly infinite)

$P = G^V$ = set of *profiles*.

Thus for $\pi \in P$,

$$\pi = (g_\alpha)_{\alpha \in V} \text{ or } \pi : V \longmapsto G.$$

Consensus method: $F : P \longmapsto G$.

Nature of Ordinal Data
(Joint project with Rudolf Wille)

View ordinal data as a relational structure
$(G; R_1, \ldots, R_n)$
G = finite set
R_i = quasiorder on G
(reflexive, transitive relation)

How can one "best" summarize
$R_1, \ldots, R_n$
by k quasiorders, where $k < n$?

What about $k = 1$?
Need a consensus theory for quasiorders.

Quasiorders on G form a lattice
NOT DISTRIBUTIVE
It is atomistic and dual atomistic.

For a profile π, will approach the construction of $F(\pi)$ by means of a "*stability*" family.

Pair (S,γ), where S is a set and $\gamma : G \longmapsto \mathcal{P}(S)$ is a one-one mapping.

Neutral consensus method F:

There is a family $\mathcal{D} = \mathcal{D}_F$ of subsets of V such that

(i) $\emptyset \notin \mathcal{D}$, but $V \in \mathcal{D}$.

(ii) $s \in \gamma(F(\pi))$ iff $s_\pi \in \mathcal{D}$,

where $s_\pi = \{\alpha : s \in \gamma(\pi(\alpha))\}$.

Goal: Relate properties of $\mathcal{D}$ with properties of γ.

Examples of stability families

1. G = a set of binary relations on a set X. (weak orders, linear orders, partial orders, quasiorders, etc.) $S = X \times X$, and

$$\gamma(R) = \{(x,y): xRy\}$$

or

$$\gamma(R) = \{(x,y): xRy \text{ fails}\}$$

2. G = finite join (or meet) semilattice.

   S = sup-irreducibles in that

$$a, b < s \Longrightarrow a \vee b < s.$$

   $\gamma(g) = \{s \in S: s \leq g\}$, or

   $\gamma(g) = \{s \in S: s \not\leq g\}$.

   (or the dual of this example)

3. Trees and Pyramids. X = set

with $|X| \geq 5$. A set $\mathcal{T}$ of subsets of X is a *tree* if

(a) $X \in \mathcal{T}$, and $\emptyset \notin \mathcal{T}$.

(b) $\{x\} \in \mathcal{T}$ for all $x \in X$.

(c) For $A, B \in \mathcal{T}$, $A \cap B \in \{\emptyset, A, B\}$

$\mathcal{T}$ is a *pyramid* if (c) is replaced by

(d) For $A, B \in \mathcal{T}$, $A \cap B \in \mathcal{T}$.

Take S = proper subsets of X, and $\gamma(\mathcal{T}) = \{A: A \in \mathcal{T}\}$. There are many similar examples.

Properties of $\mathcal{D}_F$, F neutral consensus method

Take G = Trees on set X

S = proper subsets of X (non-singleton)

$r(T) = \{C : C \in T,\ C \in S\}$

$V = \{1, 2, 3, 4, 5\}$.

1. Median consensus: $\{A : A \subseteq V,\ |A| \geq 3\}$.

$\pi = (T_1, T_2, T_3, T_4, T_5)$

$C \in F(\pi)$ iff $C \in$ at least 3 T_i's.

2. $\mathcal{D}$ not an order filter of V

$A \in \mathcal{D}$, $A \subset B$ but $B \notin \mathcal{D}$.

Just take $\{A : A \subseteq V,\ |A| = 3 \text{ or } 5\}$.

3. $A, B \in \mathcal{D}$ with $A \cap B \notin \mathcal{D}$.

Just look at #1 or #2

4. $A, B \in \mathcal{D} \Rightarrow A \cap B \neq \phi$.

True for any consensus of trees.

Example where false: $G = \mathcal{P}(X)$.

$\mathcal{D} = \{A, A', V\}$ where $A = \{1, 2, 3\}$

$\pi = (T_1, T_2, T_3, T_4, T_5)$ $T_i \subseteq X$

$F(\pi) = (T_1 \cap T_2 \cap T_3) \cup (T_4 \cap T_5)$

Some Sample Results

F = neutral consensus method with

$\mathcal{D} = \mathcal{D}_F$;

(S, γ) is a stability family.

<u>Theorem</u> 1. If $|V| \geq 2$, (a) $\Leftrightarrow$ (b)

(a) For any such F,

$$A \in \mathcal{D} \text{ implies } A' \notin \mathcal{D}.$$

(b) $\exists g, h \in G$ such that $\gamma(g) \cup \gamma(h)$ is not of the form $\gamma(v)$ for any $v \in G$.

If $|V| \geq 3$, (c) $\Leftrightarrow$ (d)

(c) For any such F,

$$B \notin \mathcal{D} \text{ implies } B' \in \mathcal{D}.$$

(d) There exist $g, h \in G$ such that $\gamma(g) \cap \gamma(h)$ is not of the form $\gamma(w)$ for any $w \in G$.

Example: $(G, \leq)$ can be a dist. lattice without $r(g) \vee r(h)$ being of the form $r(w)$

$\mathfrak{X} = \{a, b, c, d\}$

G = all weak orders containing the linear order $a < b < c < d$

Weak order: reflexive, transitive, complete

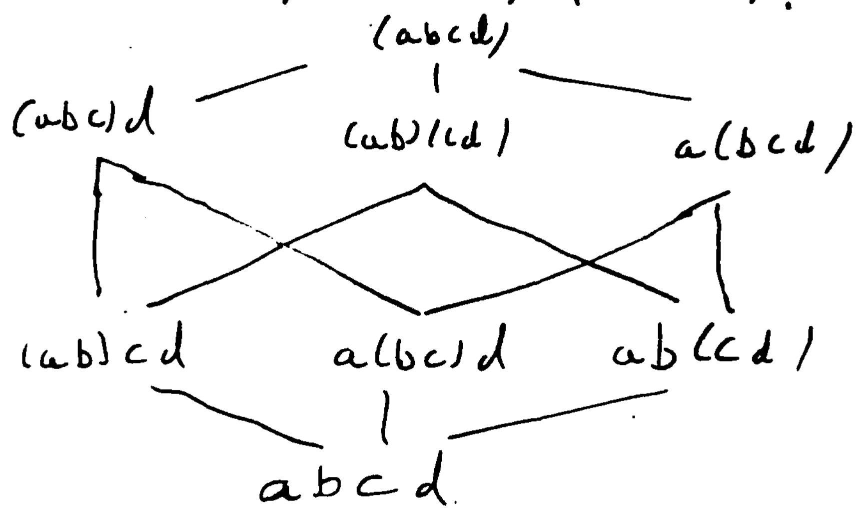
$r(w) = \{(x,y) : x \, w \, y\}$.

$G \cong 2^3$.

$$\underset{w_1}{(ab)cd} \vee \underset{w_2}{a(bc)d} = \underset{w}{(abc)d}$$

$c W a$
$c W_1 a$ fails
$c W_2 a$ fails

so $r(w_1) \cup r(w_2) \neq r(w)$.



Theorem 2. If there exist $g,h \in G$ such that

$$\gamma(g) \cup \gamma(h) \text{ is not of the form } \gamma(v),$$

but

$$\gamma(g) \cap \gamma(h) = \gamma(w) \text{ for some } w \in G,$$

then $A,B \in \mathcal{D}$ implies $A \cap B \neq \emptyset$.

Corollary. Let G be any non-distributive lattice, or any join semilattice that is not a lattice. Take S = sup-irreducibles, and

$$\gamma(g) = \{s \in S: s \leq g\}.$$

Then $A,B \in D$ implies $A \cap B \neq \emptyset$.

A Version Of Arrow's Theorem

$a_1, a_2, a_3 \in S$ have the following properties:

(1) $\exists g_1, g_2, g_3 \in G \ni (i,j,k)$ any permutation of $(1,2,3)$, then $a_i \in \gamma(g_i)$, $\{a_j, a_k\} \cap \gamma(g_i) = \emptyset$.

(2) $\exists z \in G \ni \cup_i \gamma(a_i) \subseteq \gamma(z)$.

(3) $\exists w \in G \ni \{a_1, a_2, a_3\} \cap \gamma(w) = \emptyset$.

(4) $\{a_1, a_2, a_3\}$ is *transitive* in the sense that $a_1, a_2 \in \gamma(g)$ implies that $a_3 \in \gamma(g)$.

Theorem 3. If the above conditions are satisfied and F is a neutral consensus method, then $\mathcal{D}$ is a lattice filter of V.

Remark. The above notion of transitivity of course comes from transitivity of relations:

$$xRy,\ yRz \text{ implies } xRz.$$

But what makes it of further interest is the fact that in a join semilattice, $c \le a \vee b$ makes (a,b,c) a transitive triple.

For relations, can talk about a triple $\{(a,b),(c,d),(x,y)\}$ being transitive if

$$aRb, cRd \text{ implies } xRy.$$

In general, if $T \subseteq S$ and $s \in S$, can define (T,s) to be transitive

$$\text{if } T \subseteq \gamma(g) \text{ implies } s \in \gamma(g).$$

<u>Theorem</u> 4. Let $a,b \in S$; $x,y,z \in G$.

Assume:

(1) $a \in \gamma(g) \Longrightarrow b \in \gamma(g)$.

(2) $a \in \gamma(x)$.

(3) $a \notin \gamma(y)$, $b \in \gamma(y)$.

(4) $b \notin \gamma(z)$.

Then F neutral implies $\mathcal{D}$ an order filter.

<u>Theorem</u> 5. Assume $x,y,g,z \in G$, and

(1) $\gamma(x)\cup\gamma(y) \neq \gamma(w)$.

(2) $\gamma(x)\cap\gamma(y) = \emptyset$.

(3) $\gamma(x)\cup\gamma(y) \subset \gamma(g)$.

(4) $\gamma(z) = \emptyset$.

Then F neutral implies $\mathcal{D}$ is closed under intersections.

Theorem 6. Assume $x, y, g \in G$, and

(1) $\gamma(x) \cup \gamma(y) \neq \gamma(v)$ for any v.

(2) $\gamma(x) \cap \gamma(y) = \gamma(w)$ for some w.

(3) $\gamma(x) \cup \gamma(y) \subset \gamma(g)$.

(4) $\mathcal{D}$ is an order filter.

Then $\mathcal{D}$ is a lattice filter.

Apply these and similar results to lattices having more than 1 member:

Theorem 7. G is a lattice, and S = sup-irreducibles. These are equivalent:

(1) G is not distributive.

(2) For any neutral F, $\mathcal{D}$ is a lattice filter.

Defn: A join-semilattice G is *distributive* if every principal filter of G is a distributive lattice.

Theorem 8. G is a nondistributive join semilattice not a lattice, with S its sup-irreducibles.
For any neutral consensus method F on G, $\mathcal{D}$ is an ultrafilter.

In progress: Work with distributive join-semilattices.

The setting: A stability family (S,γ) such that:
(1) $\gamma(a)\cup\gamma(b) = \gamma(v)$.
(2) $\gamma(a)\cap\gamma(b)$ is either $\gamma(w)$ or $\emptyset$.